IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nathaniel Jembere
6.30.2022

Outline

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

Executive Summary

- Summary of methodologies
  - In our analysis of Flight Launches of Flacon 9 we used scatter plots, box plots, and line graphs to visual relationship between launch site, payload, flight number, success rate, and orbit. We also used the machine learning classifiers logistic regression, KNN, decision tree, and SVM to find the best model to predict flight success.
- Summary of all results
  - The success of has increased with the number of flights
  - Higher payload have high success rates
  - Booster B4 has the highest success for large payloads
  - Decision tree classified was the best classification model to predict launch success with a 94% accuracy.

Introduction

- 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.
- If we can determine if the first stage will land, we can determine the cost of a launch

Goals:

- Predict if the Falcon 9 first stage will land successfully

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - We requested the data from SpaceX using an api. We also webscraped the data from Wikipedia using BeautifulSoup
- Perform data wrangling
  - Filtered to only Falcon 9, missing values were replaced with mean
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Compared 4 different classification models: logistic regression, KNN, decision tree, and SVM
  - Split the data into train and test sets and picked the model with the best test accuracy

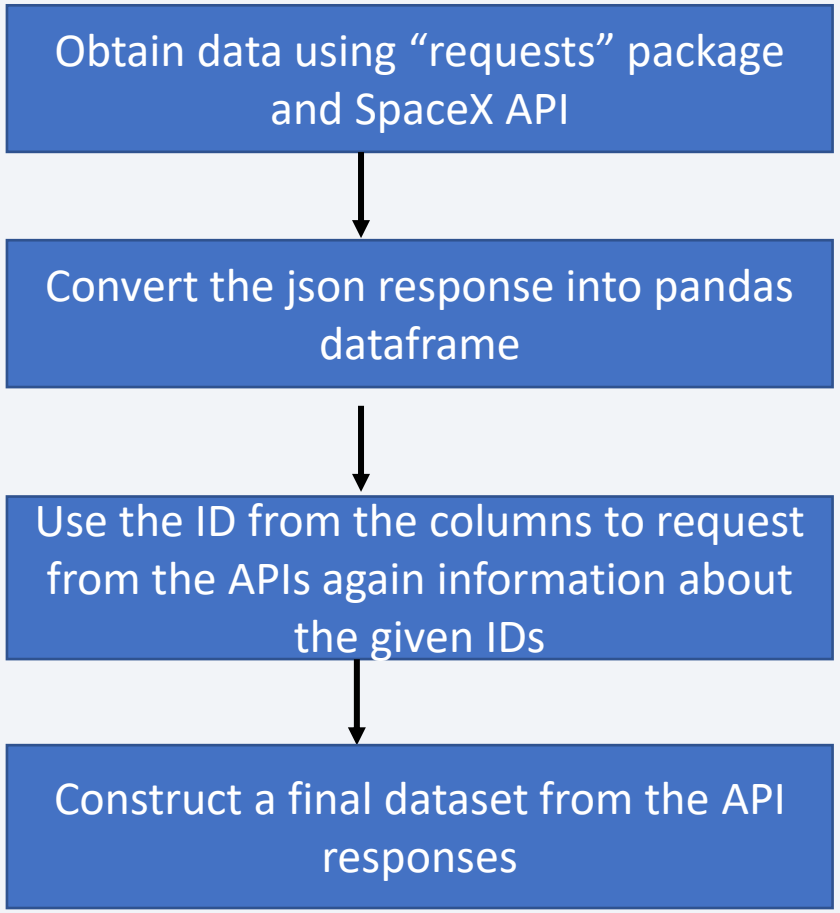
Data Collection

- We used the “request” package to pull data from the SPACEX API. The publicly available data has information such as booster versions, landing, outcomes, orbits, and much more. We filtered to Falcon 9 booster versions which are higher payloads.

- We webscraped Wikipedia page on Falcon 9 using the BeautifulSoup package for the Falcon 9 heavy packages for information on booster version, launch site, payload mass, and orbit.

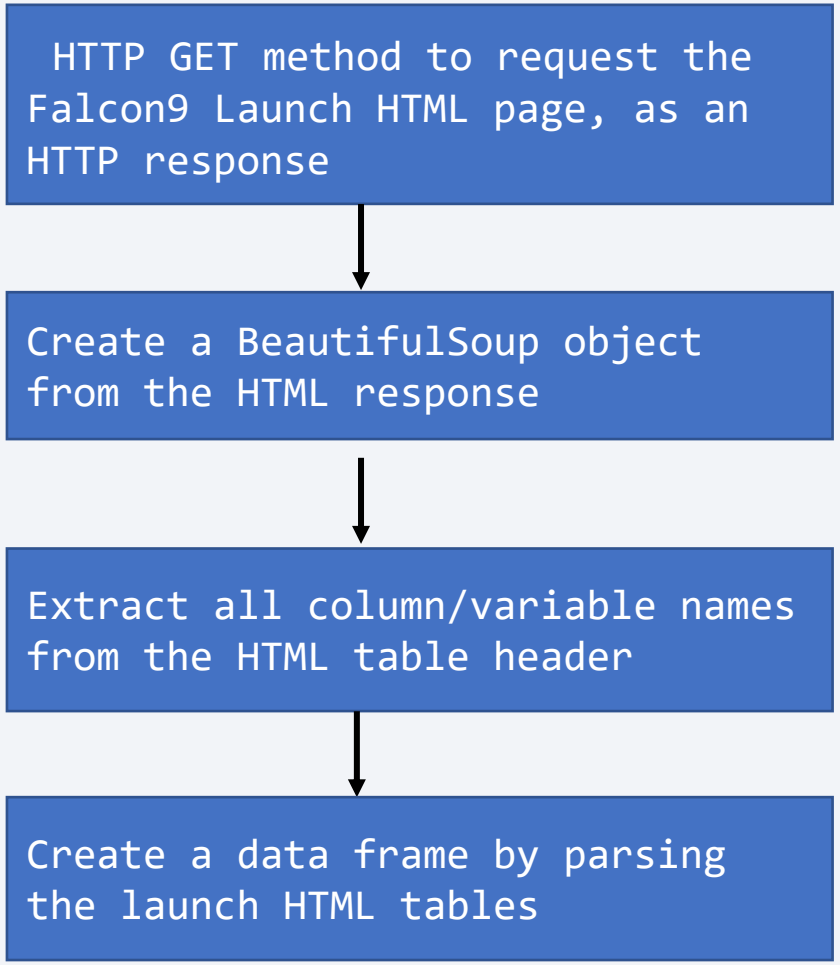
Data Collection – SpaceX API

- GitHub url: https://github.com/chefcurrywiththecode/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



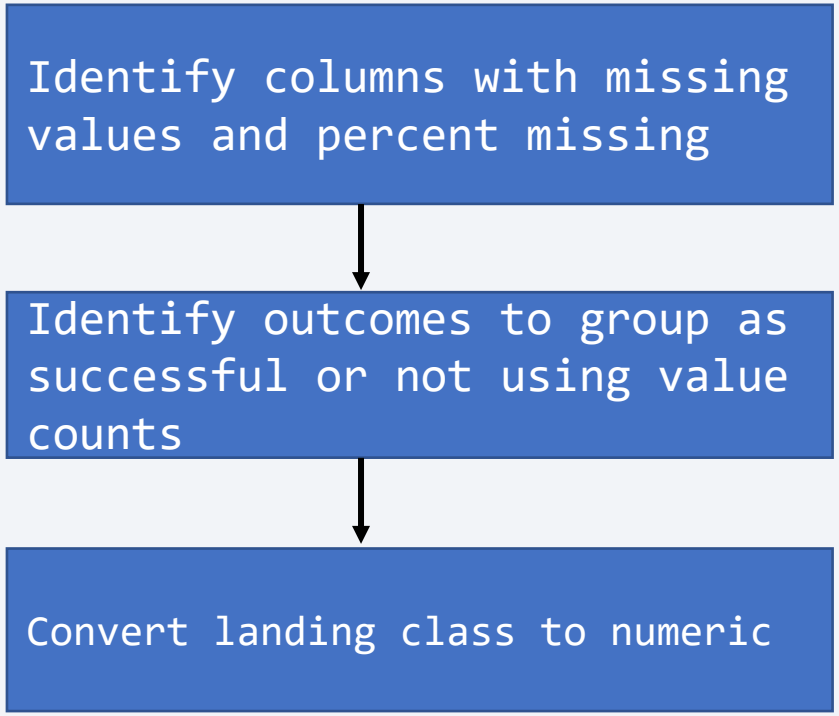
Data Collection - Scraping

- GitHub url: https://github.com/chefcurrywiththecode/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- GitHub url: https://github.com/chefcurrywiththecode/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- We used scatter plots, box plots, and line graphs to visual relationship between launch site, payload, flight number, success rate, and orbit
- Using these plots, we can gain an understanding of the relationship between two variables

- GitHub Url: https://github.com/chefcurrywiththecode/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The SQL queries performed were:
  - Selecting distinct launch sites
  - Total pay load
  - Avg mass
  - Dates of successful landing for drone ship
  - Boosters with mass between 4000 – 6000
  - Total number of missions
  - Months of successful landing outcomes in 2017

- GitHub Url: https://github.com/chefcurrywiththecode/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/jupyter-labs-eda-sql-edx_sqllite.ipynb

Build an Interactive Map with Folium

- We created the following map objects:
  - Markers: To identify launch sites
  - Circles: To show the coordinate radius around a site
  - MarkerClusters – To show the cluster of launch outcomes for a site
  - Lines- To show the distance between railway, highway, coastline, and city

- GitHub Url: https://github.com/chefcurrywiththecode/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- We added Pie and Scatterplots to our dashboard to understand the relationship between launch site, number of successful launches, and payload mass.
- The Pie chart showed the percentage of success among a site
- The scatterplot visualized the relationship between success and payload by booster and filtered by site

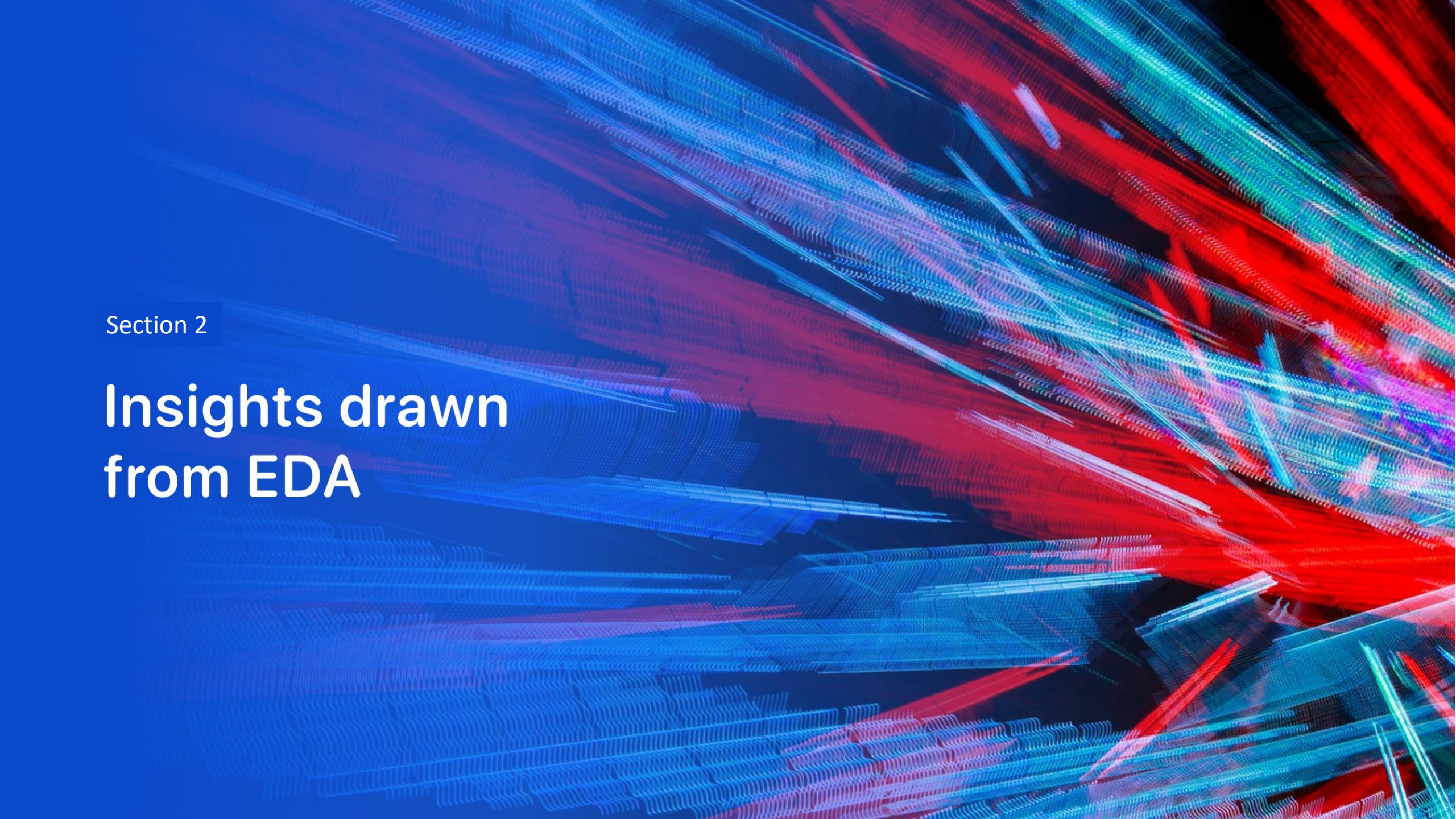
- GitHub Url: https://github.com/chefcurrywiththecode/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

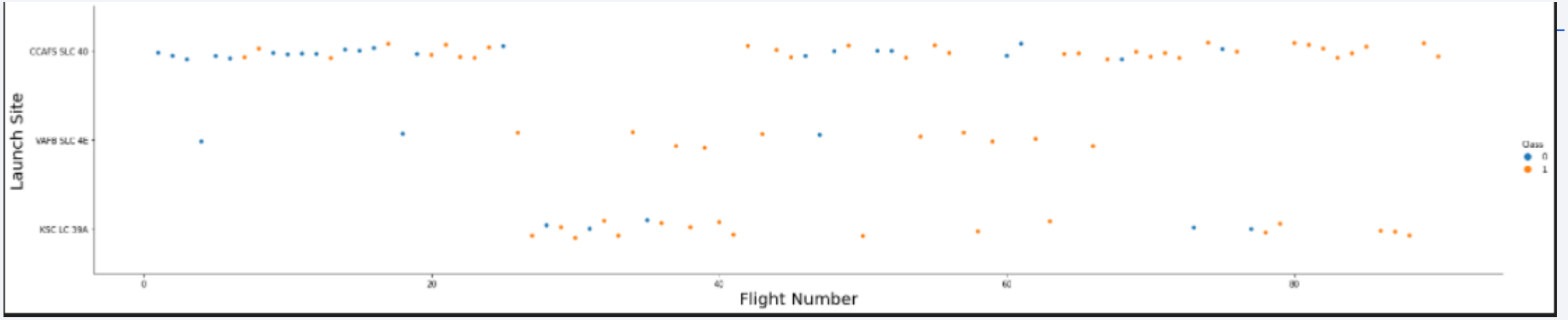
- We normalized the data using StandardScaler
- Then we split the data in to training and testing
- We used logistic regression, SVM, decision tree, and KNN to find the best training accuracy
- Gridserach was used to find the best hyperparameters for each model
- We tested the best model using testing portion to find the model with the best accuracy
- GitHub Url: https://github.com/chefcurrywiththecode/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

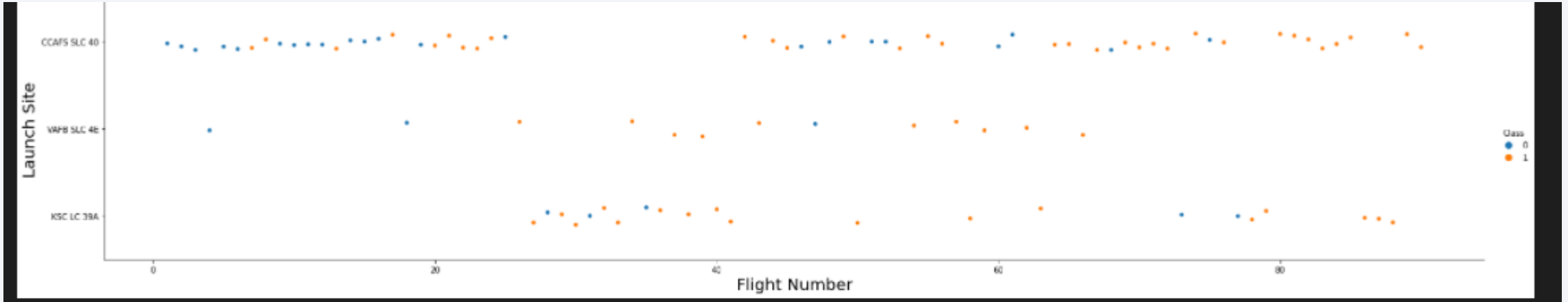
Results

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


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

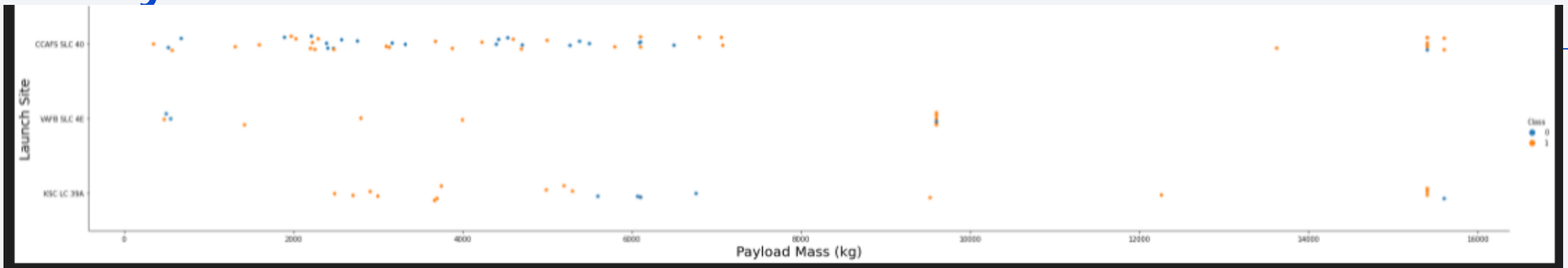


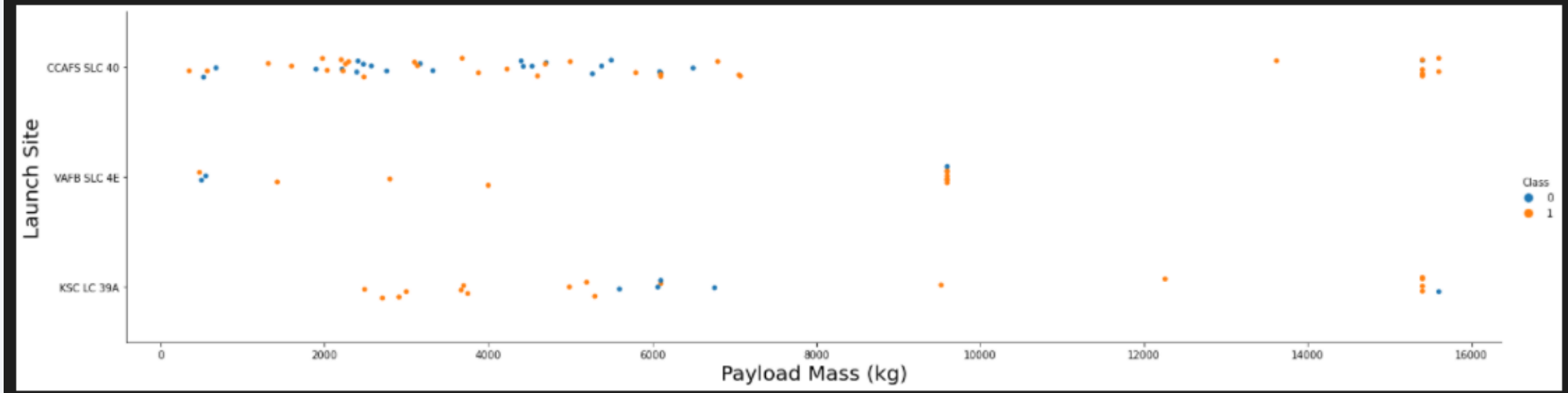


Now try to explain the patterns you found in the Flight Number vs. Launch Site scatter point plots.
Launch site CCAFS has a lot of failed launches early on. The site began having much more sucessful launches later in later flights
Launch site KSC was used the most in the middle number of fligths and the most curren launches had success.
Launch site VAFB hasn't be used much but has high success rate

Markdown

Payload vs. Launch Site

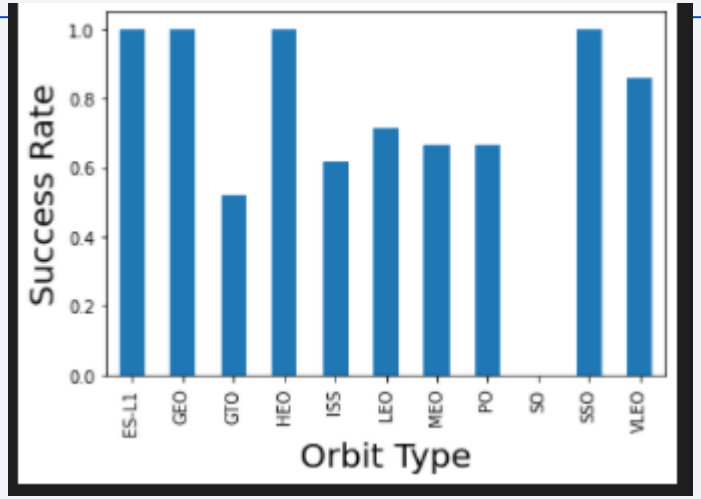


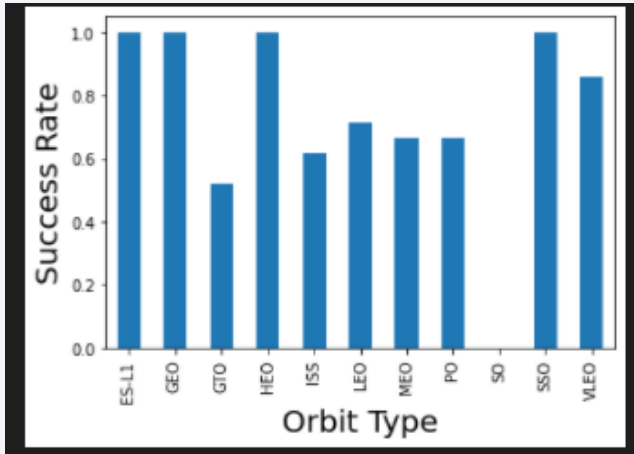


Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Heavier payloads, greater than 15000 have high success rate Payloads less than 7500 have variable succes rate for CCAFS SLC. Payloads between 5000 - 7500 have the most failure for KSC LC

Success Rate vs. Orbit Type

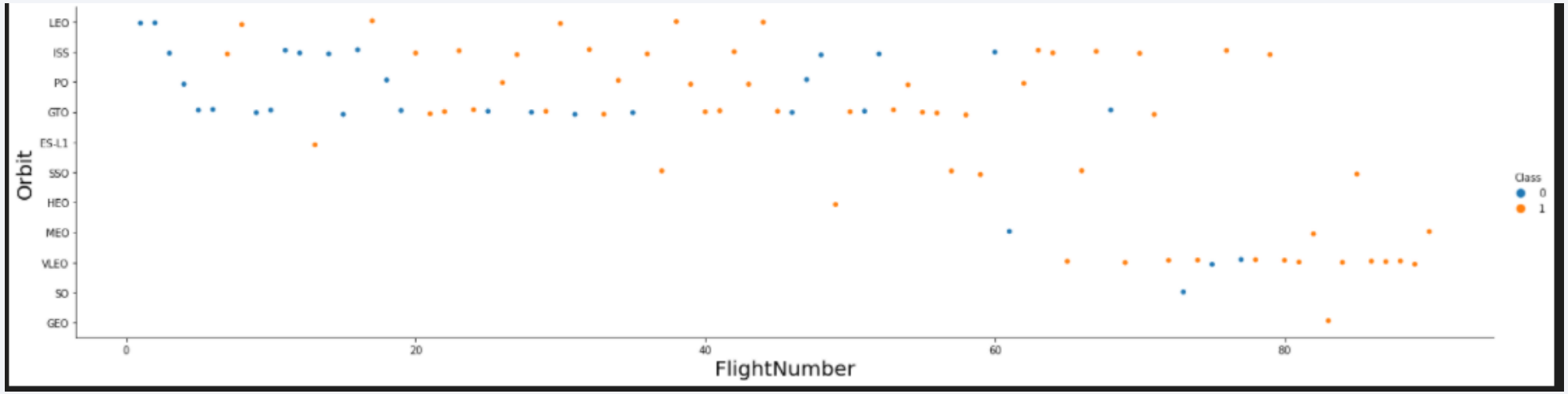


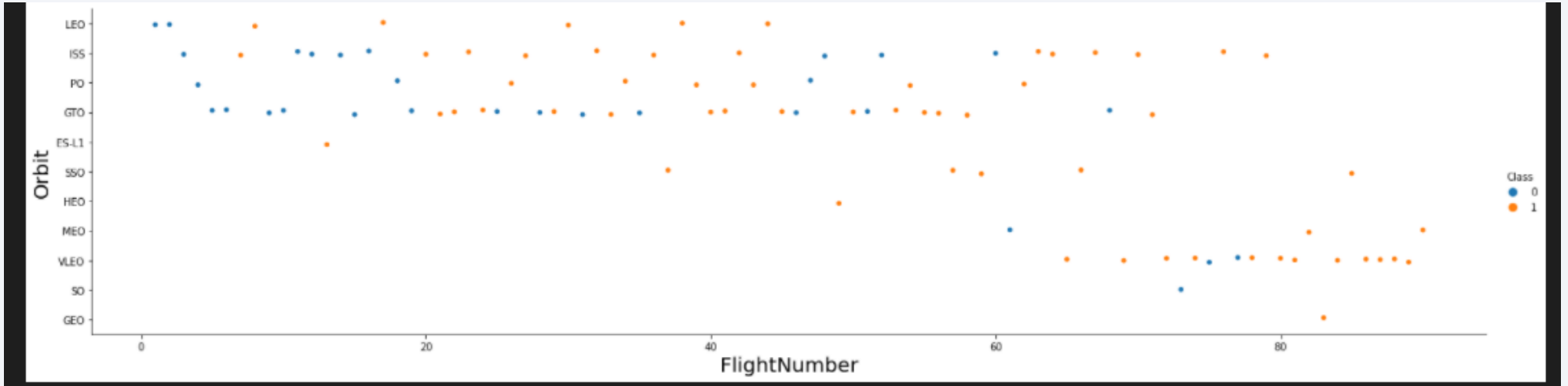


Analyze the ploted bar chart try to find which orbits have high sucess rate.
Orbit SO has a success rate of 0. Orbit GTO has the second lowest sucess rate of 0.5. ESL1, GEO, ISS, and SSO have the highest success rate of 1.

Markdown

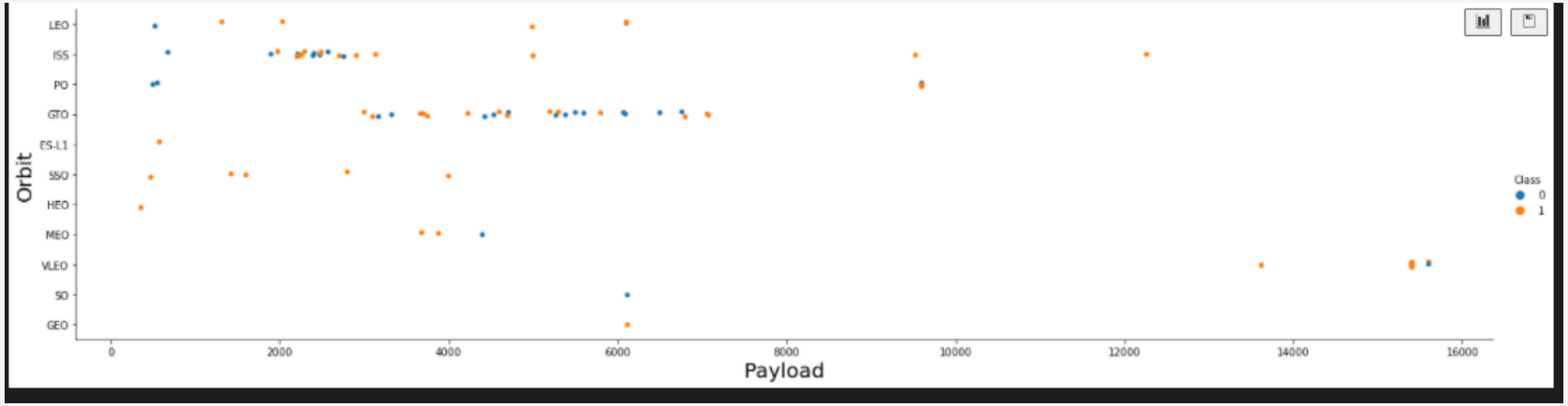
Flight Number vs. Orbit Type

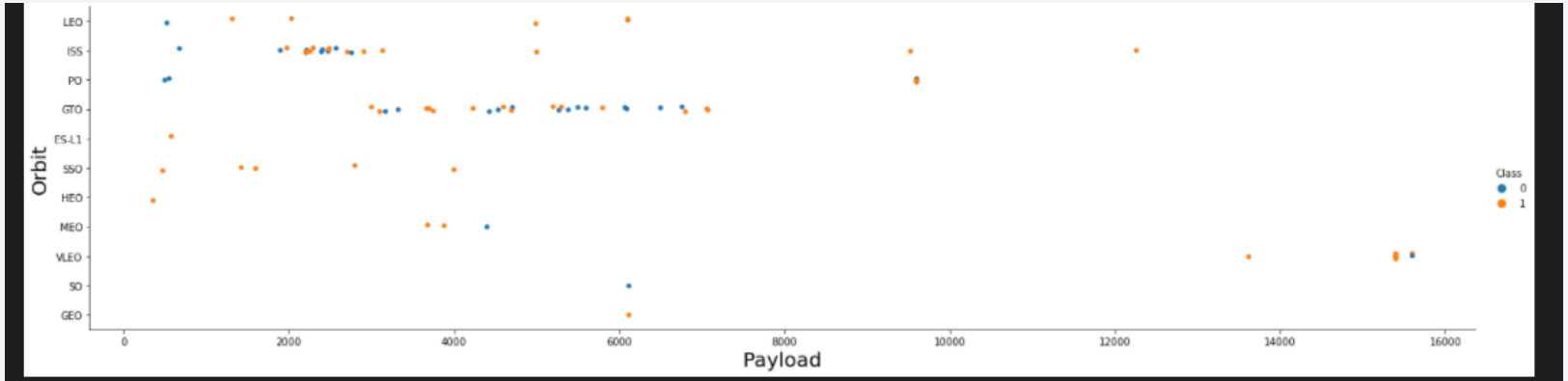




You should see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit. VLEO has had a lot flights later on which have been successful. ISS had a lot of failed flights early on.

Payload vs. Orbit Type





With heavy payloads the successful landing or positive landing rate are more for Polar,VLEO and ISS.

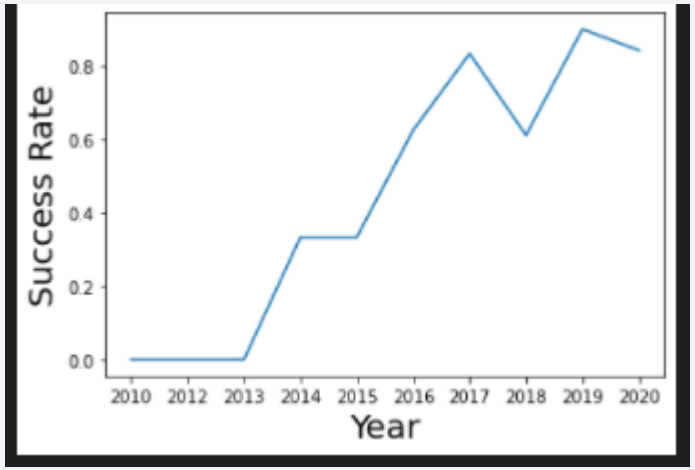
However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

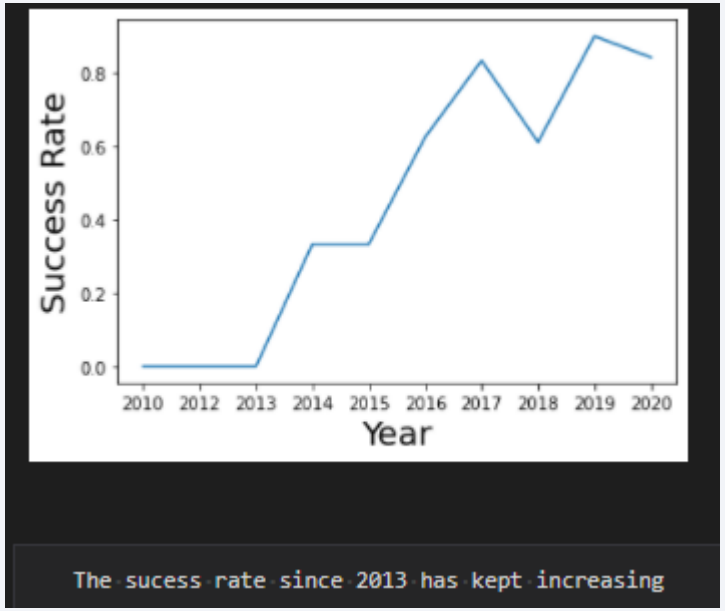
SSO has a lot of success with the lighter payloads.

A lot of the lighter payloads under 2000 kg had failures

Markdown

Launch Success Yearly Trend





All Launch Site Names

- The unique launch sites are: CCAFS SLC 40, CCAFS LC 40, KSC LC 39A, and VAFB SLC 4E

```
%sql select distinct LAUNCH_SITE from SPACEXTBL
```

✓ 0.1s Python

```
* sqlite:///my_data1.db
Done.
```

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

- The query above selects and outputs unique values of Launch Site

Launch Site Names Begin with 'KSC'

```
%sql select * from SPACEXTBL where LAUNCH_SITE like 'KSC%' limit 5
```

✓ 0.1s

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
16-03-2017	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
15-05-2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

- The query looked for launch sites having KSC using the function like
- Only the first 5 results were outputted using limit

Total Payload Mass

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER = 'NASA (CRS)'
 ✓ 0.3s
 * sqlite:///my_data1.db
Done.

sum(PAYLOAD_MASS__KG_)
                 45596
```

- Summing the total pay load mass in the dataset, we find a total payload mass of 45,596 kg

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION = 'F9 v1.1'
```

✓ 0.9s

```
 * sqlite:///my_data1.db
Done.
```

avg(PAYLOAD_MASS__KG_)
2928.4

- The avg payload mass for F9 v1.1 was 2928.4 kg

First Successful Ground Landing Date

```
%sql select min(date) from SPACEXTBL where `Landing_Outcome` = 'Success (drone ship)'
✓ 0.5s

* sqlite:///my_data1.db
Done.

min(date)
06-05-2016
```

- The earliest successful drone ship landing was on May 5th, 2016

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:
  - F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

```
%sql select BOOSTER_VERSION from SPACEXTBL where `Landing _Outcome`='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

✓ 0.4s

```
* sqlite:///my_data1.db
Done.
```

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

- There was 4 successful drone ship landings with payload mass between 4000 and 600. They were: 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) as count from SPACEXTBL
```

✓ 0.8s

```
 * sqlite:///my_data1.db
Done.
```

count
101

- There was a total of 101 successful and failed mission outcomes.

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass are:
  - F9 B5 B1048.4, F9 B5 B1048.5, F9 B5 B1049.4, F9 B5 B1049.5 , F9 B5 B1049.7, F9 B5 B1051.3, F9 B5 B1051.4, F9 B5 B1051.6, F9 B5 B1056.4, F9 B5 B1058.3, F9 B5 B1060.2, F9 B5 B1060.3,

```
%sql select distinct BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)

✓ 0.1s
 * sqlite:///my_data1.db
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

- There were 12 boosters which carried the maximum payload.

2017 Launch Records

```
%sql select SUBSTR(DATE, 4, 2) AS 'MONTH',`Landing _Outcome`,Booster_Version,LAUNCH_SITE from SPACEXTBL where SUBSTR(DATE, 7, 4) = '2017' and `Landing _Outcome` = 'Success (ground pad)'
```

✓ 0.1s Python

```
* sqlite:///my_data1.db
Done.
```

MONTH	Landing _Outcome	Booster_Version	Launch_Site
02	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
05	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
06	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
08	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
09	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
12	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

- There were 6 launches in 2017 where there was successful landing on ground pad. They were in the months, Feb, May, June, Aug, Sept, and Dec.

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

```
%%sql select `Landing _Outcome`, count(`Landing _Outcome`) as count from SPACEXTBL where date between '04-06-2010' and '20-03-2017' and `Landing _Outcome` like '%Success%' group by `Landing _O
```

49] ✓ 0.8s Python

```
 * sqlite:///my_data1.db
Done.
```

Landing _Outcome	count
Success	20
Success (drone ship)	8
Success (ground pad)	6

- There were between 2010-06-04 and 2017-03-202 there were 20 successes, 8 drone ship success, and 6 ground pad success.

Section 3

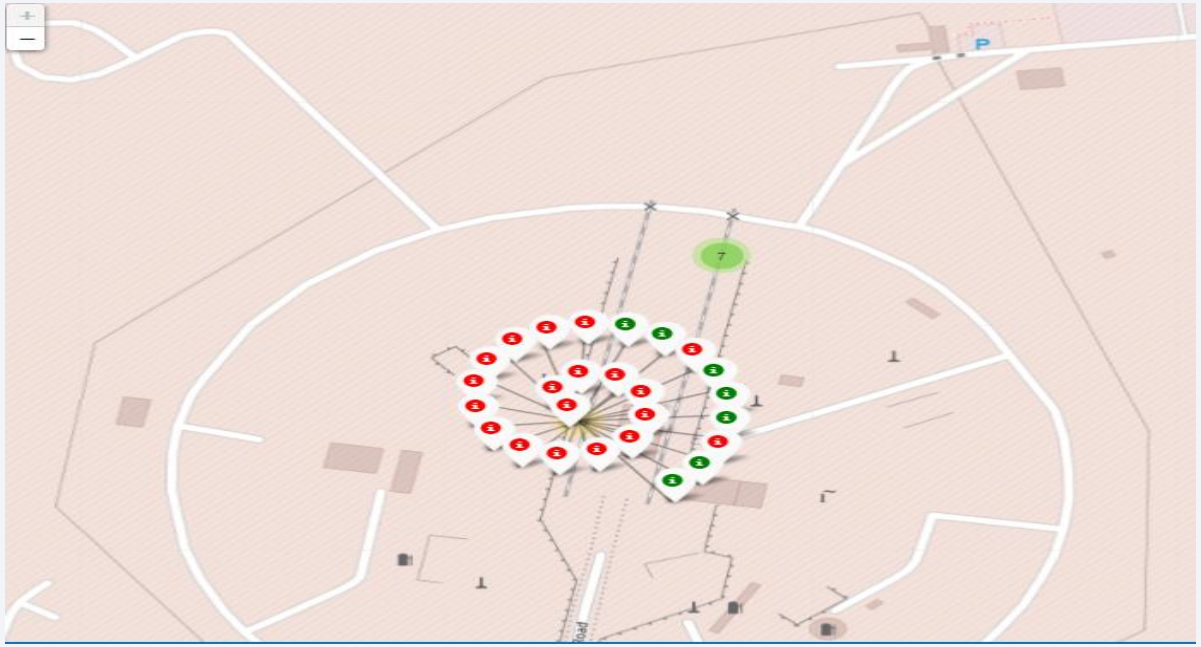
Launch Sites Proximities Analysis

Space X Launch Site Locations

- All launch sites are in the US, with two being on the east coast and two being on the west coast.

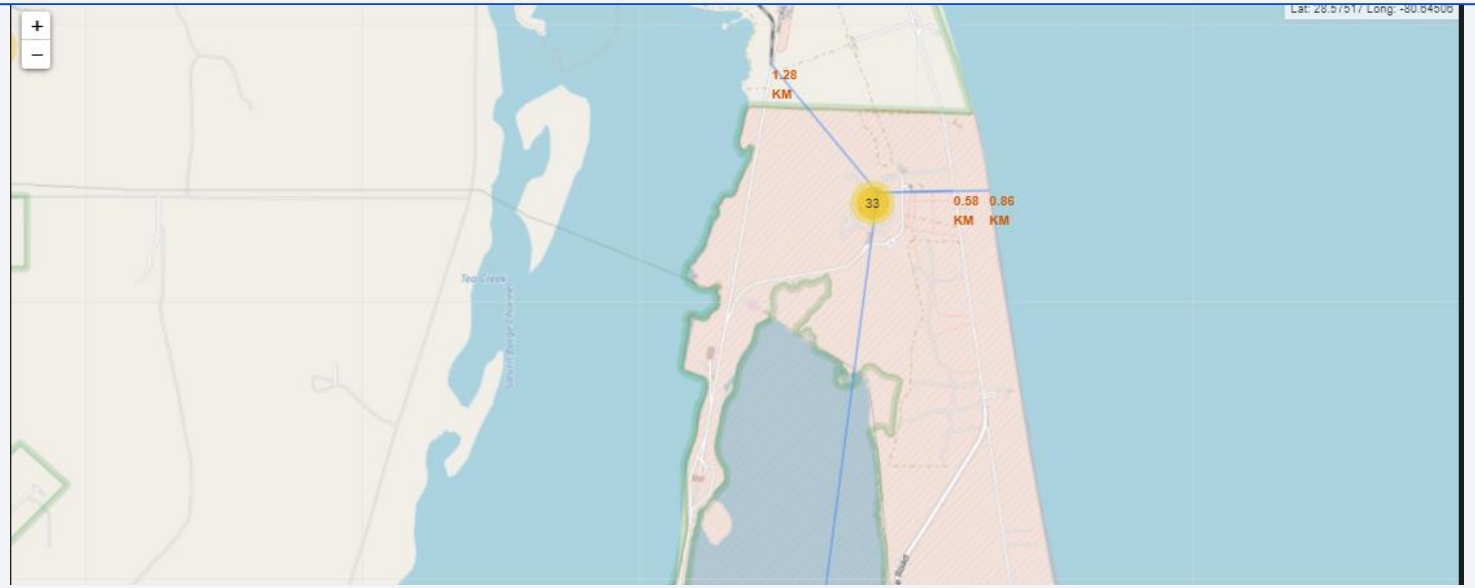
- The launch sites which are on the same coast are close to each other

Rate of launch success for a launch site

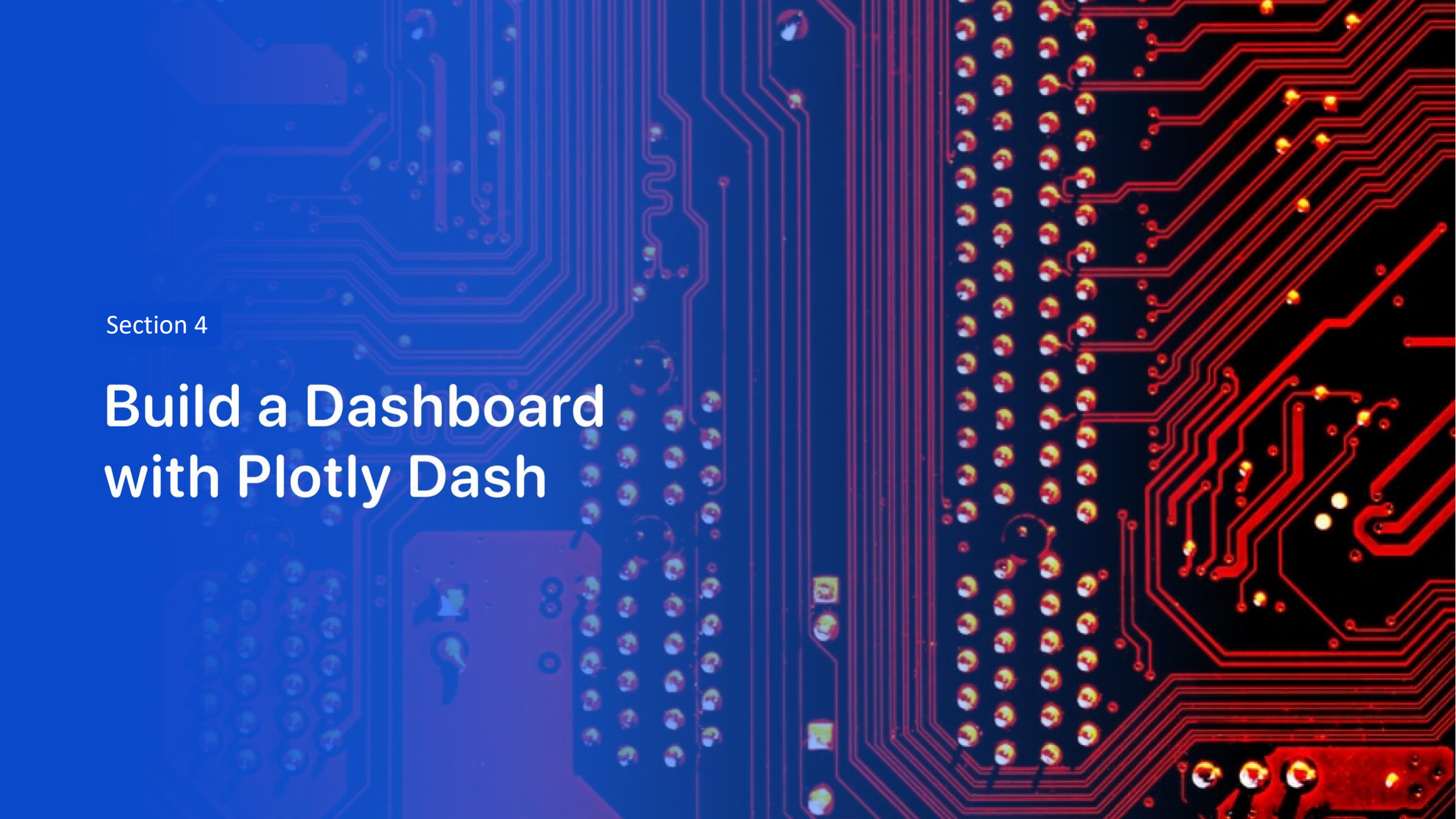


- The successful launches are in green while the unsuccessful launches are in red
- This launch site had more unsuccessful than successful launches

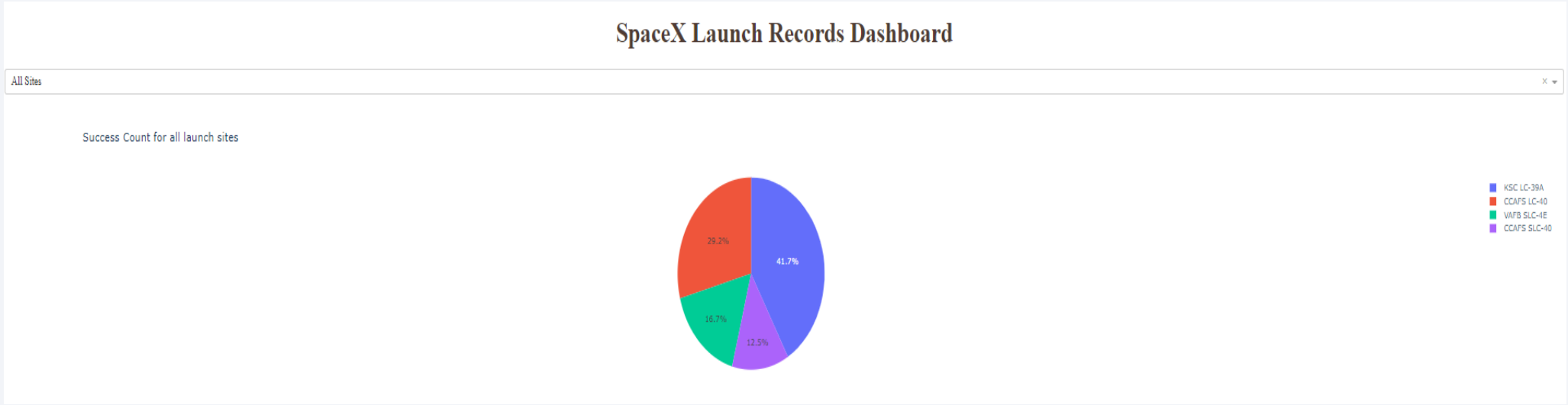
Launch Sites Are Far From Cities



- Distances from coastline, highways, railways, and cities are captured by the blue line
- The further landmark are cities from the launch site. The closest is highways and coastline

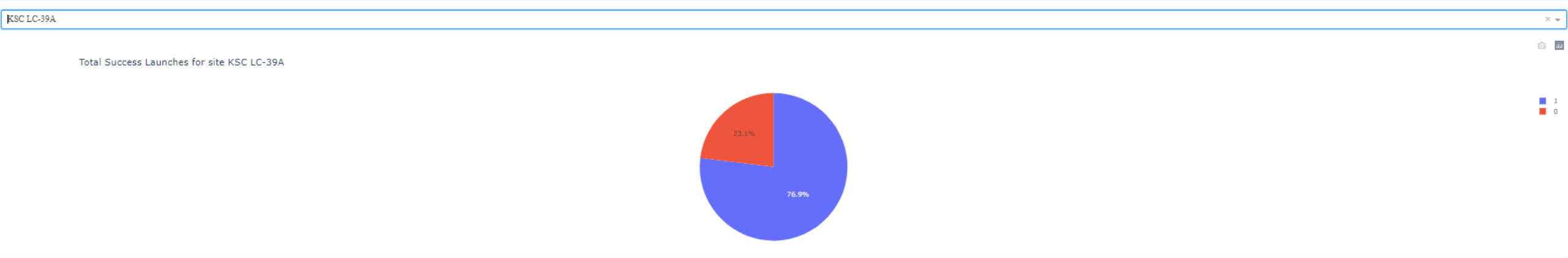

Section 4
Build a Dashboard with Plotly Dash

KSC LC-39A has the highest number of launch success



- Each color represents a launch site
- The size of the pie is proportional to the how many of the overall launch success was attributed to the site
- The site KSC LC-39A has the highest number of launch success (in blue)

KSC LC-39A has the highest launch success ratio



- Successful launch are in blue and unsuccessful launches are in red
- KSC LC-39A has the highest launch success ratio with a success ratio of 76.9%

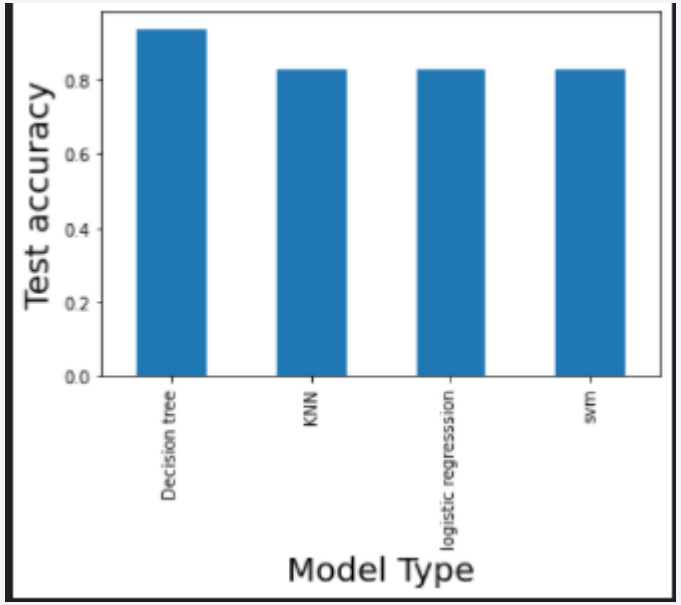
Booster B4 has the highest success for large payloads

- In this scatterplot, the y axis is success and the x axis is payload mass
- The payload mass has been filtered to between 4000 kg and 9000kg
- Among the higher payloads, only three boosters have been used. The booster with the highest success rate is B4.

Section 5

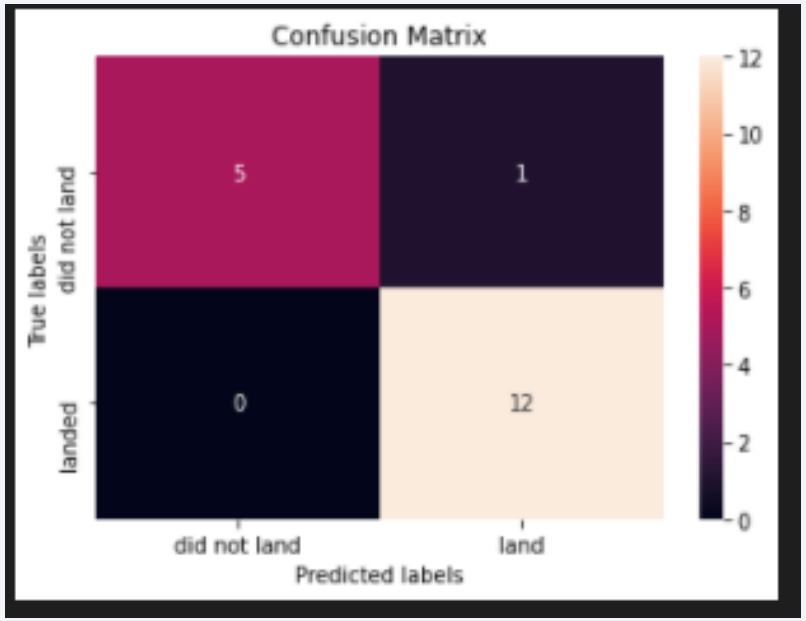
Predictive Analysis (Classification)

Classification Accuracy



- KNN, Logistic Regression and SVM all had the same test accuracy of 0.83. Decision tree had the highest accuracy with 0.94.

Confusion Matrix



- This is the confusion matrix for the decision tree. We can see true and predicted lands matched on 12 rows and did not land matched 5 times. There was only 1 prediction error.

Conclusions

- The number of successful landings has increased with the number of flights
- KSC LC-39A is the best site to launch a flight from
- The best classifier for prediction landing outcomes was the decision tree model
- The decision tree accuracy on the test data was 94.4%
- We can use the decision tree model to predict the landing of the next flight based on the flight details

Appendix

- Git hub database: https://github.com/chefcurrywiththecode/Data-Science-and-Machine-Learning-Capstone-Project

Thank you!