IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - EDA with SQL and Visualizations
  - Building Interactive maps
  - Producing a dashboard
  - Classifying using predictive analysis
- Summary of all results
  - EDA was used to determine the relationships between Successful and failed launches
  - Machine learning was used to find the predictors to find optimal conditions for a successful launch

Introduction

- We are a competitor of Space X. We are using their open-source data to develop a model to determine if a launch would likely be successful given the conditions at launch time.
- We want to analyze different launch conditions to be able to determine the best conditions to launch a rocket into space successfully. This will help to cut down on R&D and production costs.

Section 1

Methodology

Methodology

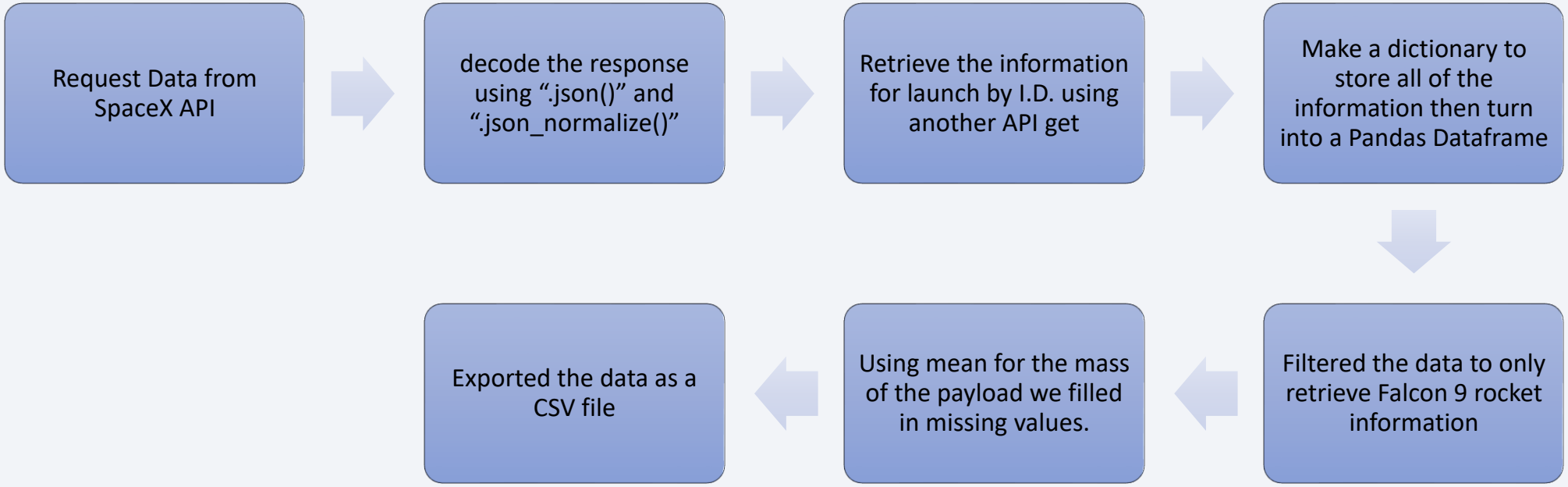
Executive Summary

- Data collection methodology:
  - We collected data in two manners:
    - Accessing SpaceX Rest API
    - Web Scraping Wikipedia
- Perform data wrangling
  - Filtered, dealt with missing values, and used one-hot encoding to make it useful for ML
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Built, tuned , and evaluated multiple models to determine the best models.

Data Collection

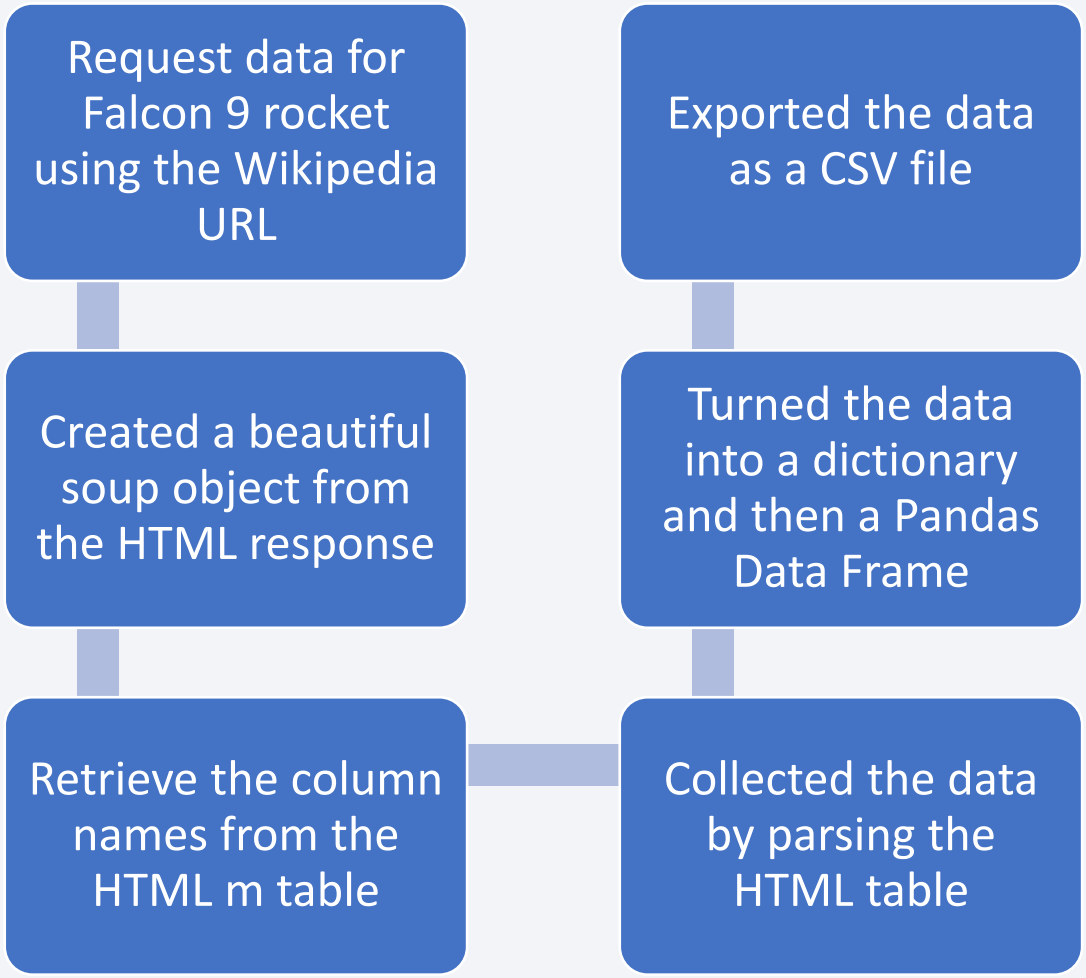
- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API



API Github file

Data Collection - Scraping

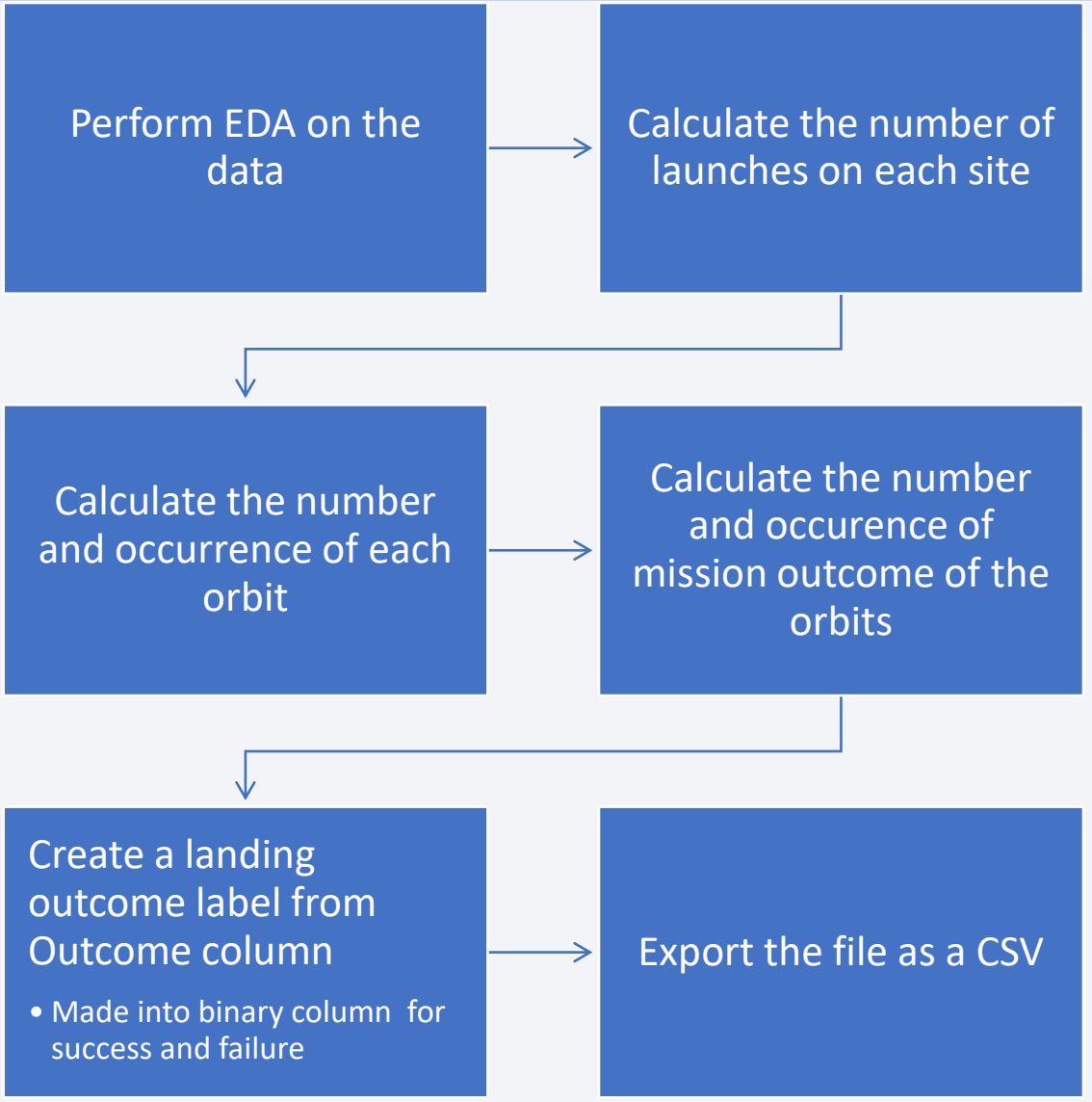


- Web scraping Github

Data Wrangling

In this data we encountered data that was well classified for the purposes of reporting for SpaceX, however it is not optimal for our final goal of creating a ML model. Successes and failures were recorded for differing manners of each. We were able to explore the data and then consolidate it into one binary class for success and failure to make it ML friendly.

[Data wrangling Github]



EDA with Data Visualization

- We used a Bar graph to show the success rate of each orbit
- A line graph to show change in success rate of launches per year
- We used Scatterplots to :
  - Show the relationship between :
    - launch number and Payload mass
    - launch number and Launch Site
    - Launch Site and Payload Mass
    - Flight number and Orbit
    - Payload Mass and Orbit

Data Viz EDA Github

EDA with SQL

- Using SQL queries we:
  - Displayed each unique launch site
  - Displayed 5 records where launch sites begin with the string 'CCA'
  - Displayed total payload mass carried by boosters launched by NASA (CRS)
  - Displayed average payload mass carried by booster version F9 v1.1
  - Displayed date when the first succesful landing outcome in ground pad was acheived
  - Displayed names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Displayed total number of successful and failure mission outcomes
  - Displayed names of the booster versions which have carried the maximum payload mass using a subquery
  - Displayed records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
  - Displayed count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

SQL EDA Github

Build an Interactive Map with Folium

- Marked all launch sites:
  - We marked and labeled each launch site as to determine latitude and longitude. This will allow us to determine if location plays an impact in the success of a launch
- Marked successes and failures:
  - We added a popout for each launch. We then colored them green if the launch was a success and red if it failed. This allows us to visualize if the launch site had an impact on the success of the launch
- Calculate the distances between a launch site to its proximities:
  - We added measurements to the nearest Railway, City, Highway, and Coastline

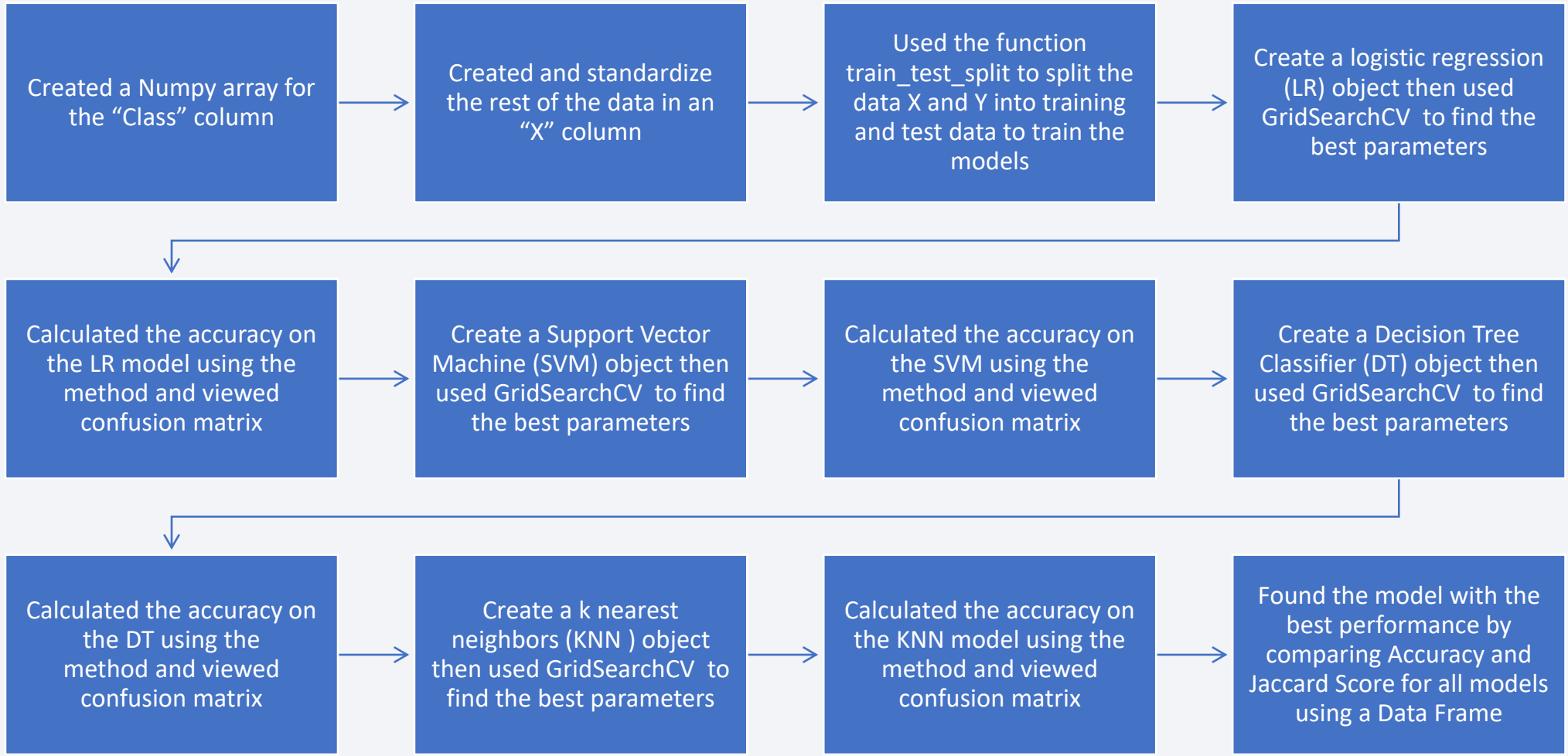
Folium Github

Build a Dashboard with Plotly Dash

- To our dashboard we added:
  - Launch site dropdown to select by launch site
  - A pie chart to show successful vs. failed launches by site
  - A slider to allow filtering by payload size
  - A scatterplot to show relationship between Payload and success rate for the booster sizes

SQL EDA Github

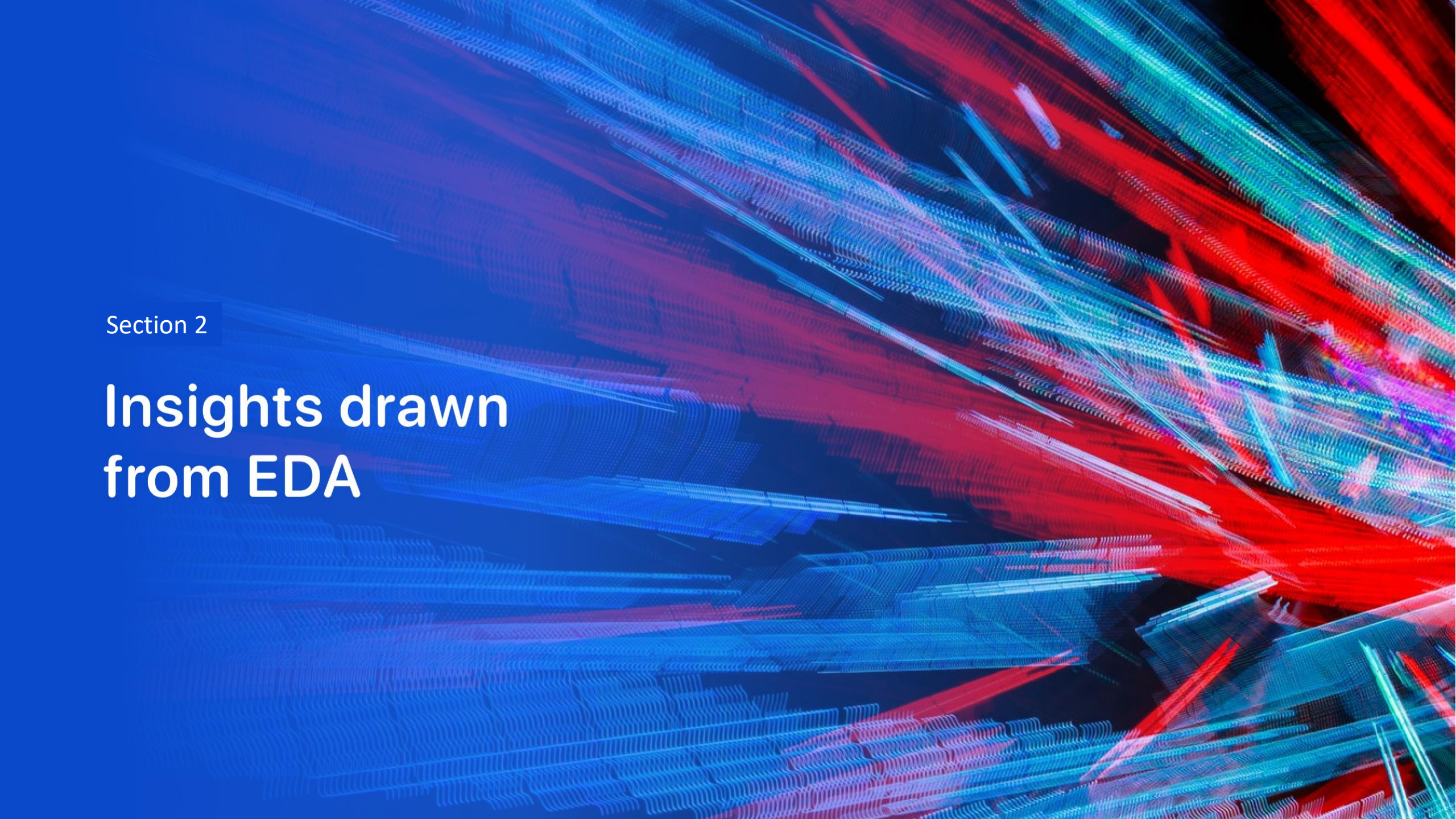
Predictive Analysis (Classification)



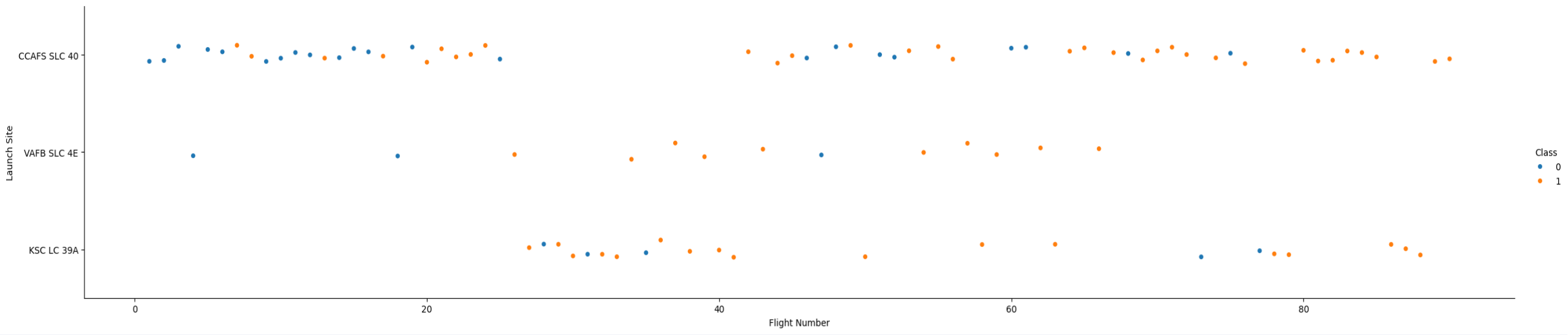
ML Github

Results

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


Section 2
Insights drawn from EDA

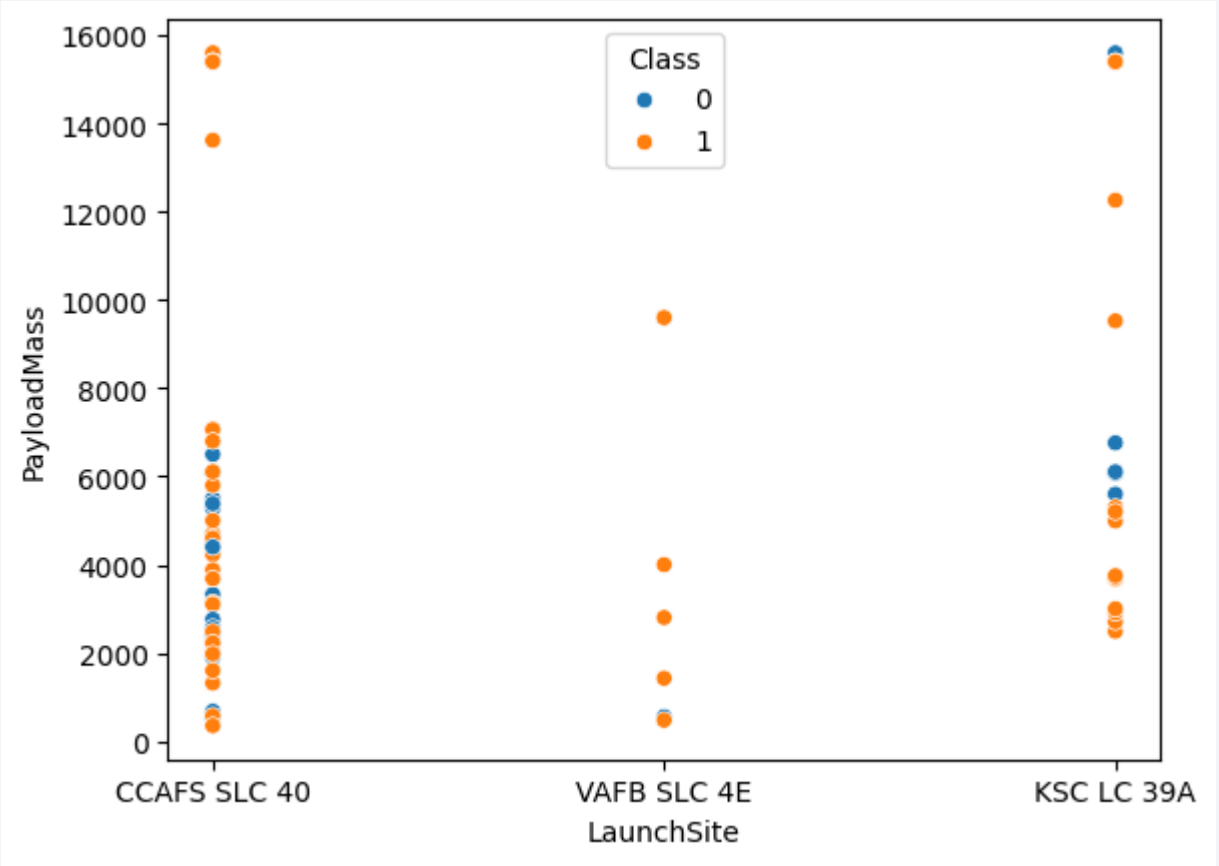
Flight Number vs. Launch Site



- This grap shows:
  - The earlies flights (1-6) all failed but that is likely due to a new process.
  - CCAFS SLC 40 is the most used site for launches
  - The more flights that were launched seems to increase the likelihood of success as is to be expected

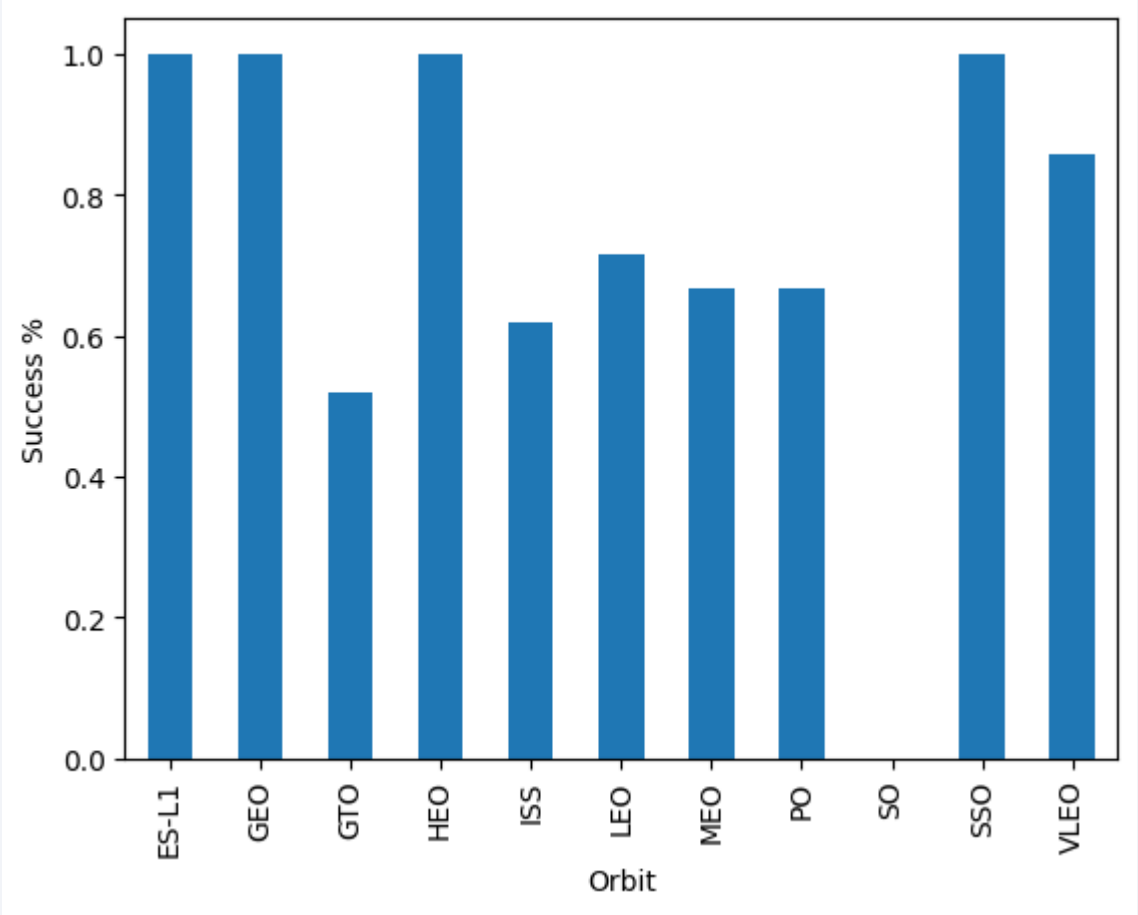
Payload vs. Launch Site

- This scatterplot shows that:
  - Maajority of launches occurred had a mass under 10,000
  - VAFB SLC has not had a launch over 10,000



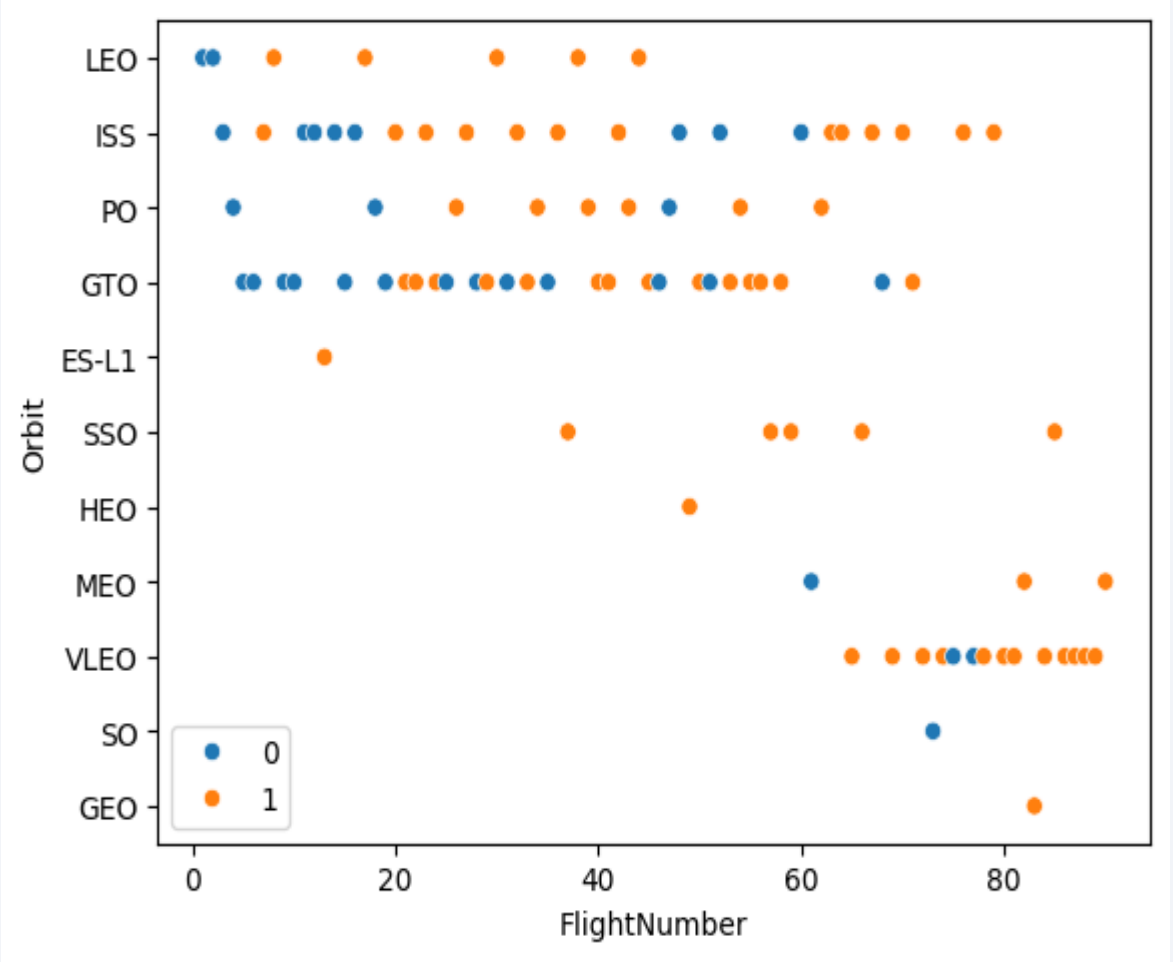
Success Rate vs. Orbit Type

- This bar chart shows:
  - ES-L1, GEO, HEO, and SSSO have a 100%B success rate
  - SO orbit has a 0% success rate. (Although there may have been no attempt to reach that orbit)
  - The other Orbits appear to have at least a 50% success



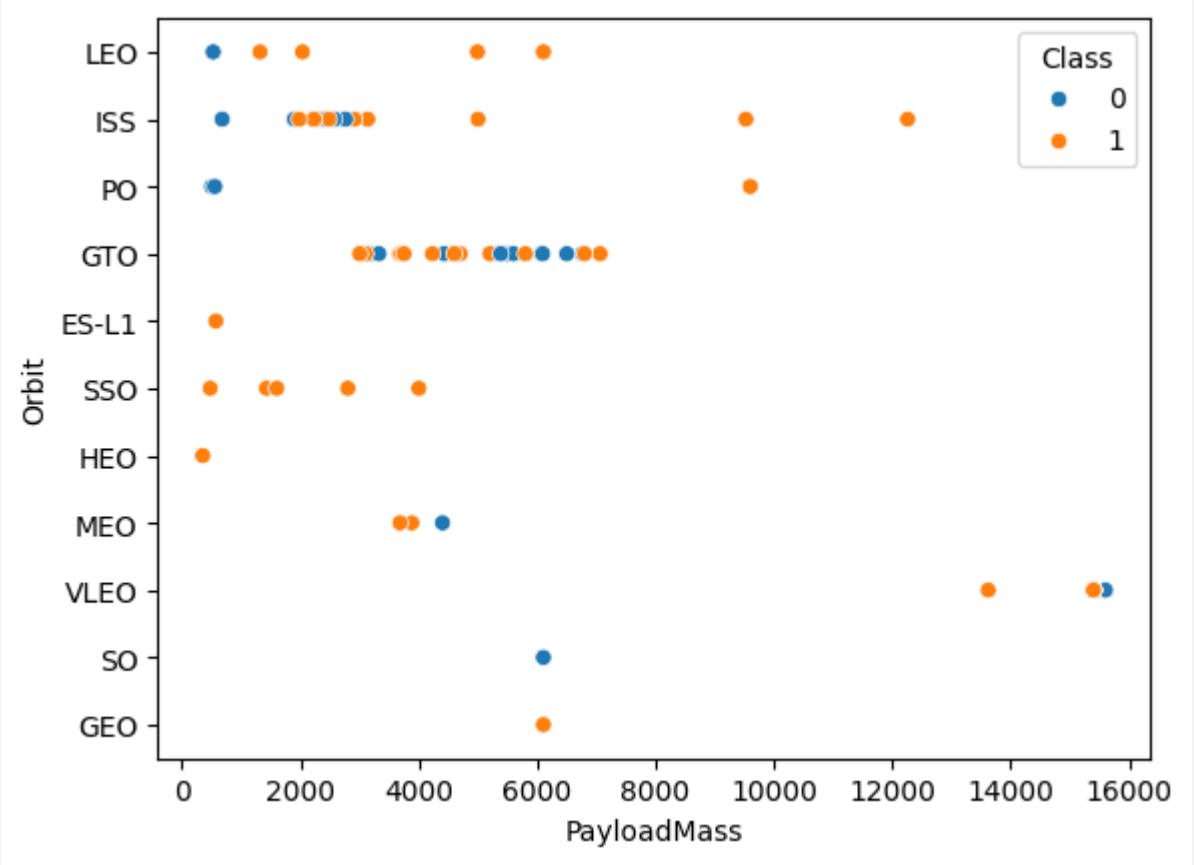
Flight Number vs. Orbit Type

- The earliest launches failed regardless of the orbit
- For LEO, ISS, and PO it appears that success came after learning from failures of the first few launches



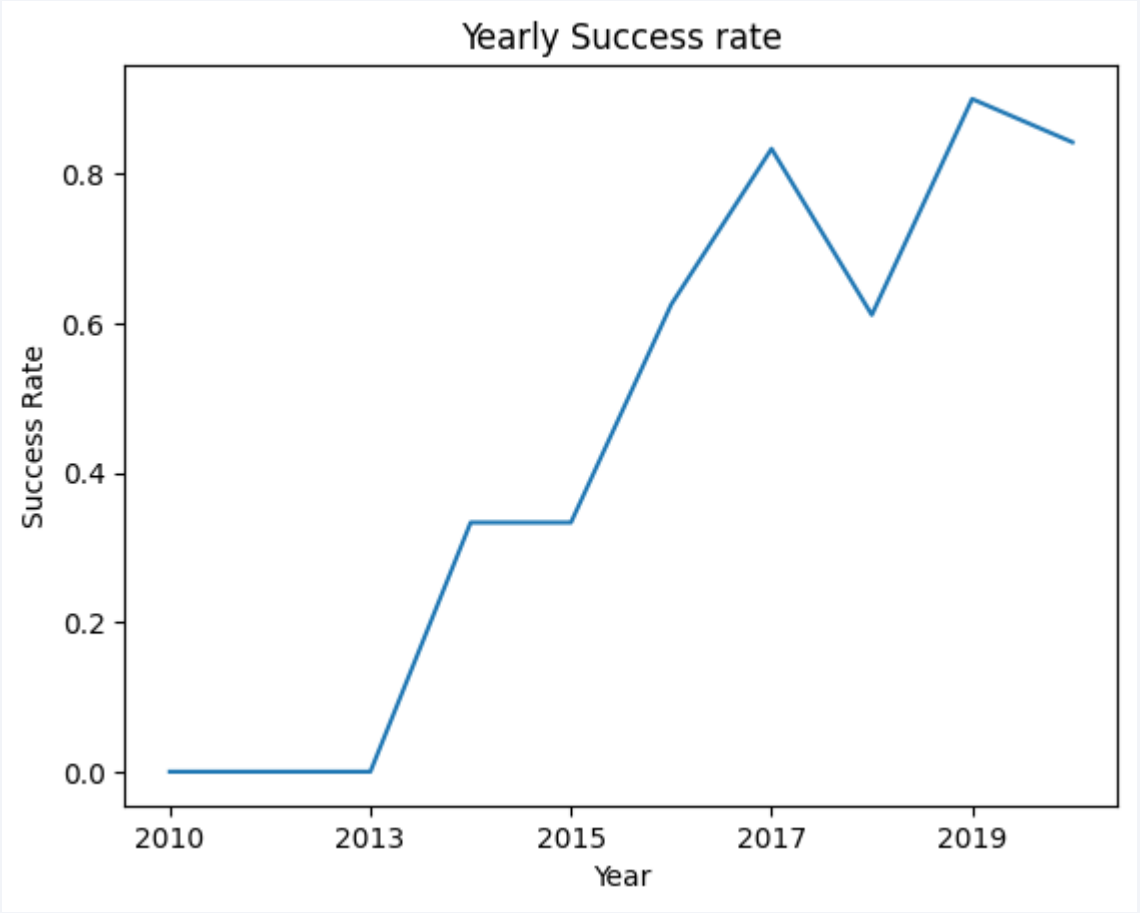
Payload vs. Orbit Type

- Generally as Payload increased the success rate increased.
- The heavier payloads succeeded more often
- GTO appears to be less successful as payload increases



Launch Success Yearly Trend

- The early years had no successes as is to be expected with a new product
- As the years went on they had more successes
- 2018 was the outlier in the continued growth for launch success being the only year of decrease from the year prior



All Launch Site Names

```
%%sql

select distinct Launch_Site from SPACEXTABLE
```

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

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

- This are all of the unique launch sites used by SpaceX

Launch Site Names Begin with 'CCA'

```
%%sql

select * from spacextbl where Launch_Site LIKE 'CCA%' limit 5;
```

```
 * sqlite:///my_data1.db
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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

- These are the first five launches from a site starting with “CCA”

Total Payload Mass

```
%%sql

select sum(PAYLOAD_MASS__KG_) from spacextbl where Customer = 'NASA (CRS)'
```

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

sum(PAYLOAD_MASS__KG_)
45596

- The total Mass carried by SpaceX for NASA is 45,596kg

Average Payload Mass by F9 v1.1

```
%%sql

select avg(PAYLOAD_MASS__KG_) from spacextbl where Booster_Version LIKE 'F9 v1.1';
```

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

avg(PAYLOAD_MASS__KG_)
2928.4

- The average payload carried by the Falcon 9 is 2,928.4kg

First Successful Ground Landing Date

```
%%sql

select min(Date) as min_date from spacextbl where Landing_Outcome = 'Success (ground pad)';
```

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

min_date
2015-12-22

- The first successful ground pad landing was December 12 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
select Booster_Version from spacextbl where (PAYLOAD_MASS__KG_> 4000 and PAYLOAD_MASS__KG_ < 6000)
and (Landing_Outcome = 'Success (drone ship)');
```

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

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

- F9 FT B1022, F9 FT B1026, F9 FT B1021.2, and F9 FT B1021.2 are the boosters the had a paylkoad between 4,000 kg and 6,000kg to successfully land on a drone ship

Total Number of Successful and Failure Mission Outcomes

```
%%sql

select Mission_Outcome, count(Mission_Outcome) as counts from spacextbl group by Mission_Outcome;
```

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

Mission_Outcome	counts
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- In total they classified 100 as successes and 1 as a failure

Boosters Carried Maximum Payload

```
%%sql

select Booster_Version, PAYLOAD_MASS__KG_ from spacextbl where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from spacextbl);
```

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

- These are the boosters that have carried the maximum payload of 15,600kg

2015 Launch Records

```
%%sql

select Landing_Outcome, Booster_Version, Launch_Site from spacextbl where Landing_Outcome = 'Failure (drone ship)' and strftime('%Y', Date) = '2015'
```

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

Landing_Outcome	Booster_Version	Launch_Site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- These are the two launches that failed in 2015

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

```
%%sql

select Landing_Outcome, count(*) as LandingCounts from spacextbl where Date between '2010-06-04' and '2017-03-20'
group by Landing_Outcome
order by count(*) desc;
```

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

Landing_Outcome	LandingCounts
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

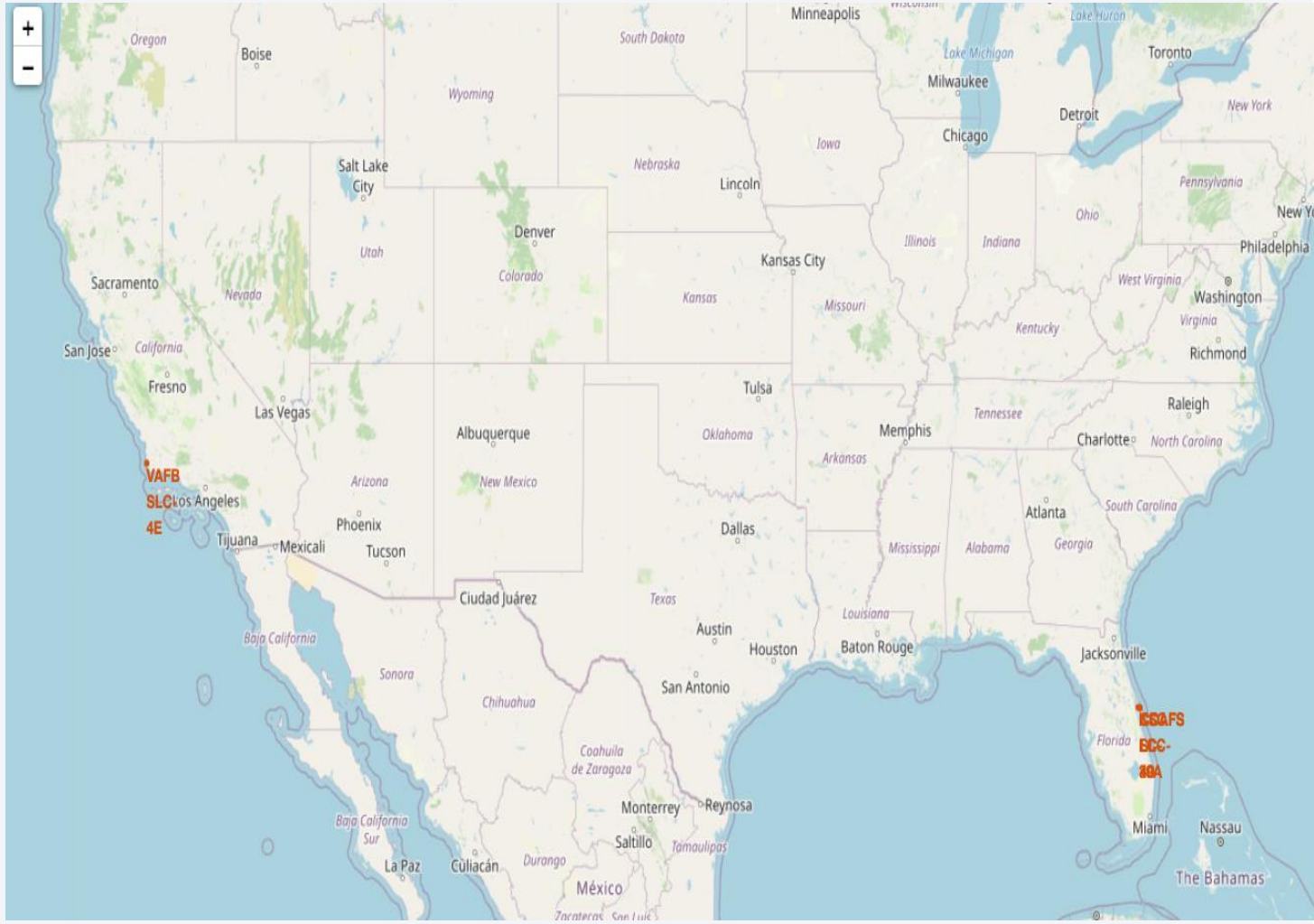
- Between 2010-06-04 and 2017-03-20 these are the outcomes that occurred in descending order.

Section 3

Launch Sites Proximities Analysis

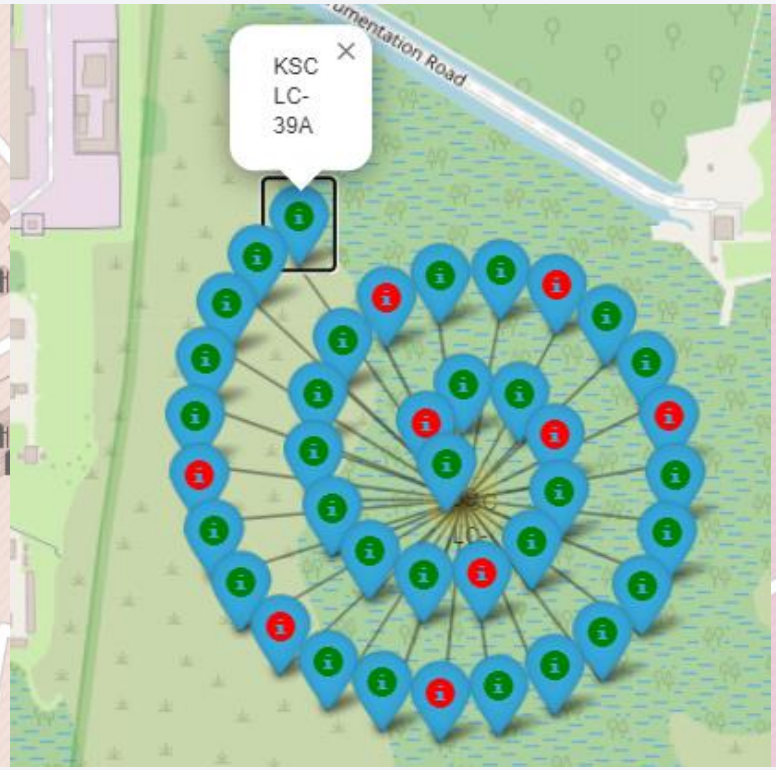
Launch locations on folium

- The launch locations are on opposite coasts of the U.S.
- Florida is near the NASA launch site.
- Florida is closer to the equator which in theory makes it a more optimal launch location.



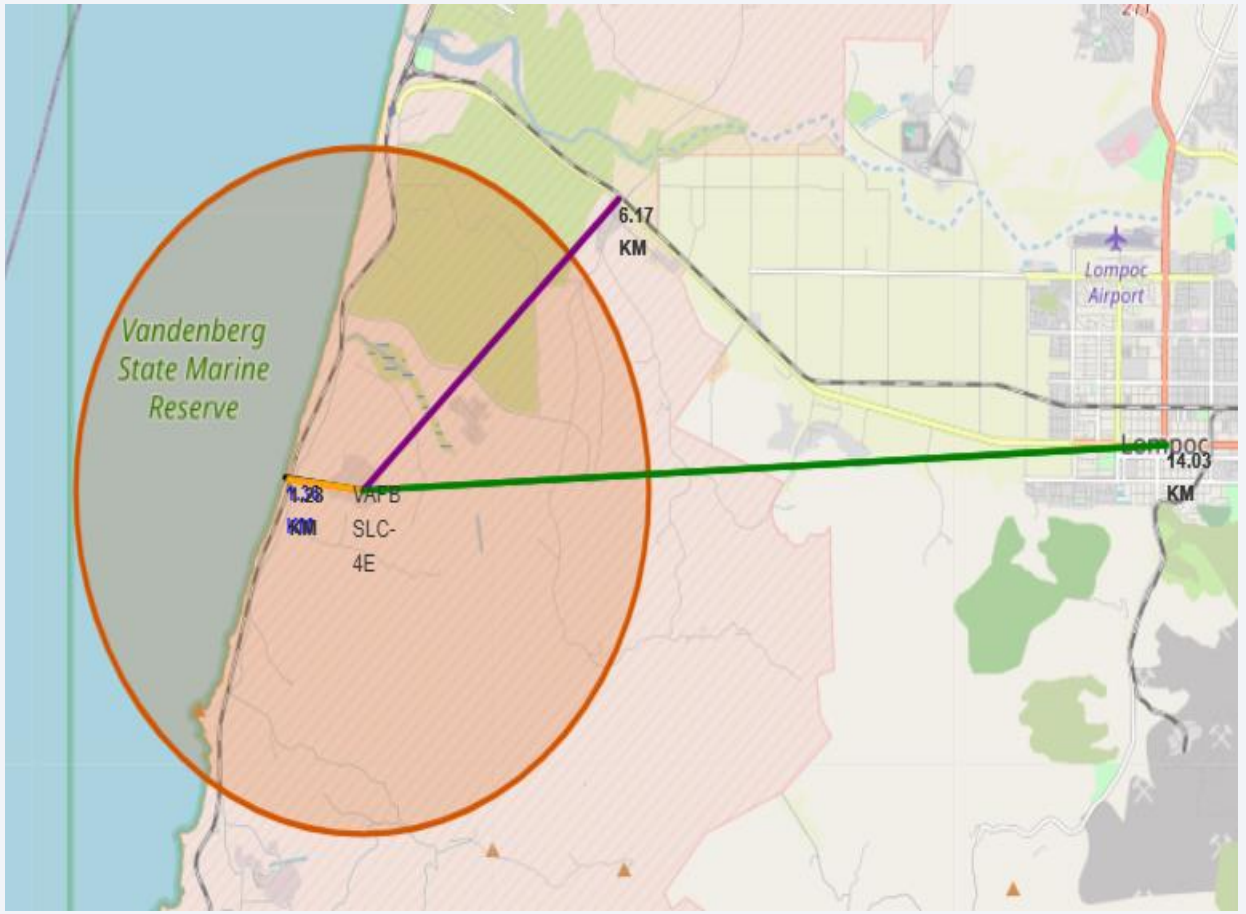
Launch outcome popouts

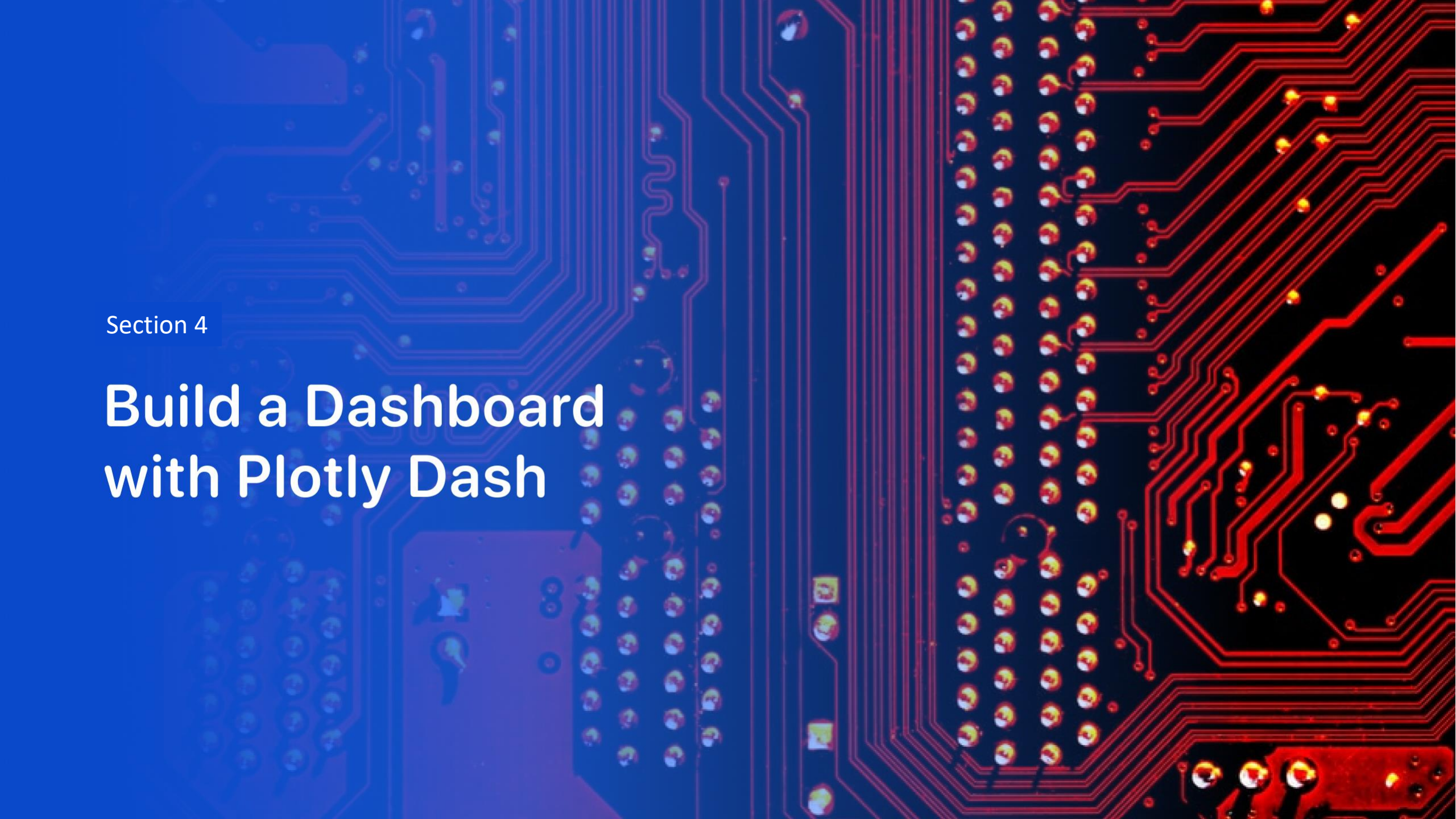
- This shows successful launches in green and Failed in red for the different launch locations.



Distance in km from nearest points

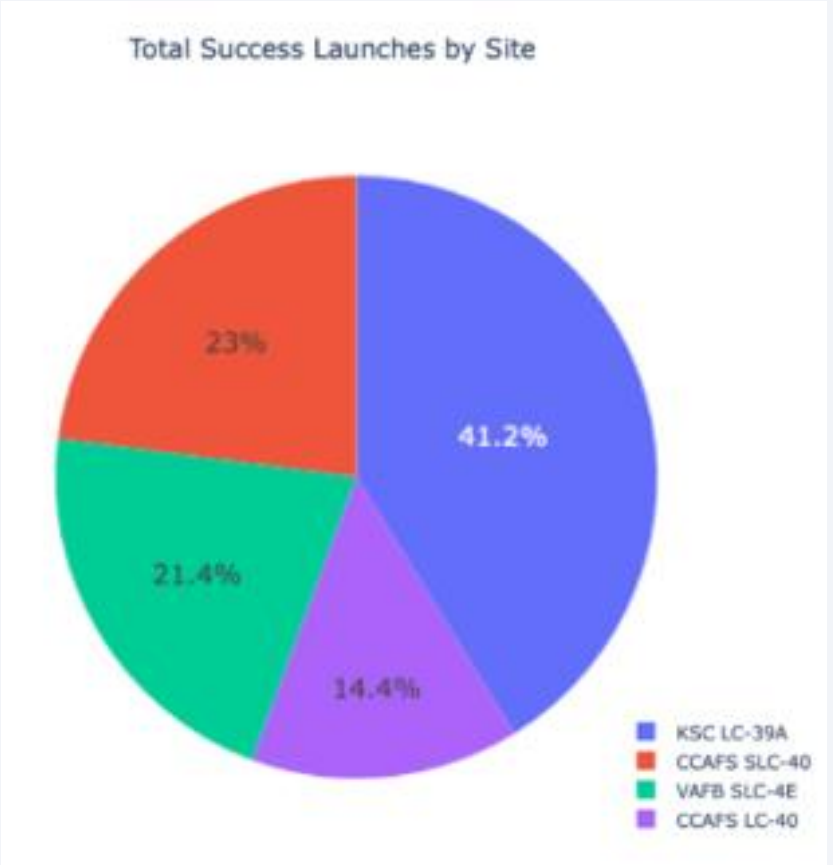
- VAFB SLC-4E is roughly:
  - 14.03km from the nearest city
  - 6.17km from the nearest highway
  - 1.28km from the nearest railway
  - 1.36km from the nearest coast




Section 4
Build a Dashboard with Plotly Dash

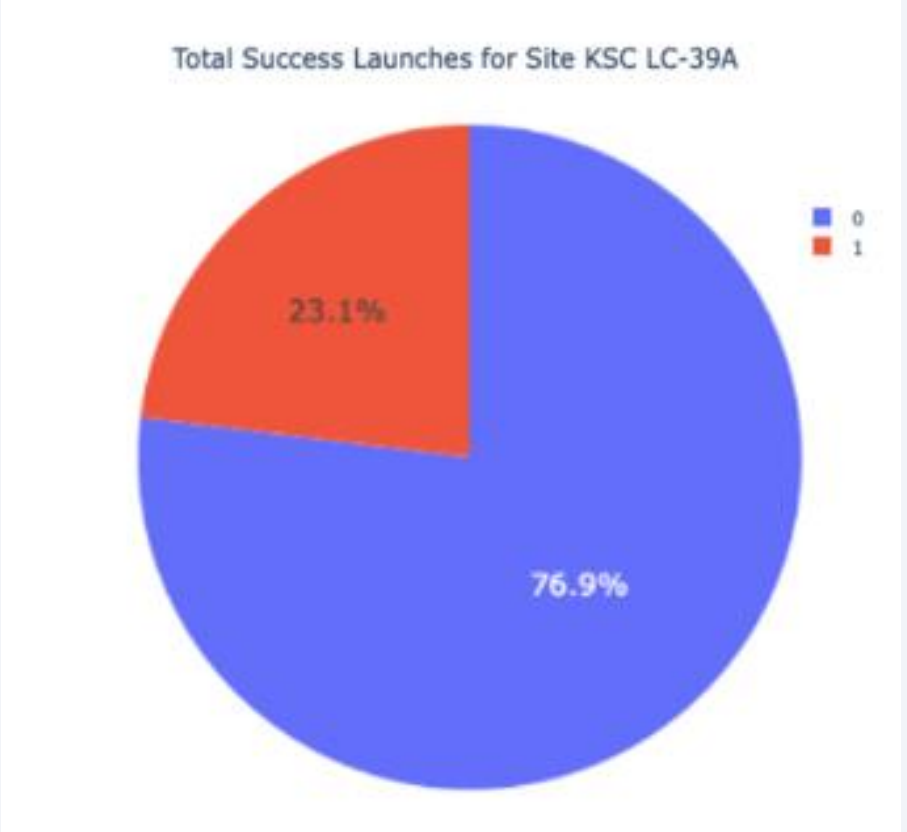
Percentage of successful launches by location

- Site KSC LC-39A has the most launches that ended successfully
- The rest are very similar in total successful launches with CCAFS LC-40 being slightly lower.



Success rate of KSC LC-39A

- Site KSC LC-39A has success rate of 76.9%
- The failure rate is 23.1%



<Dashboard Screenshot 3>

- This shows that a higher success rate for the upper half of the payload scale
  - This is a little misleading as there are more launches at this weight portion

Section 5

Predictive Analysis (Classification)

Classification Accuracy

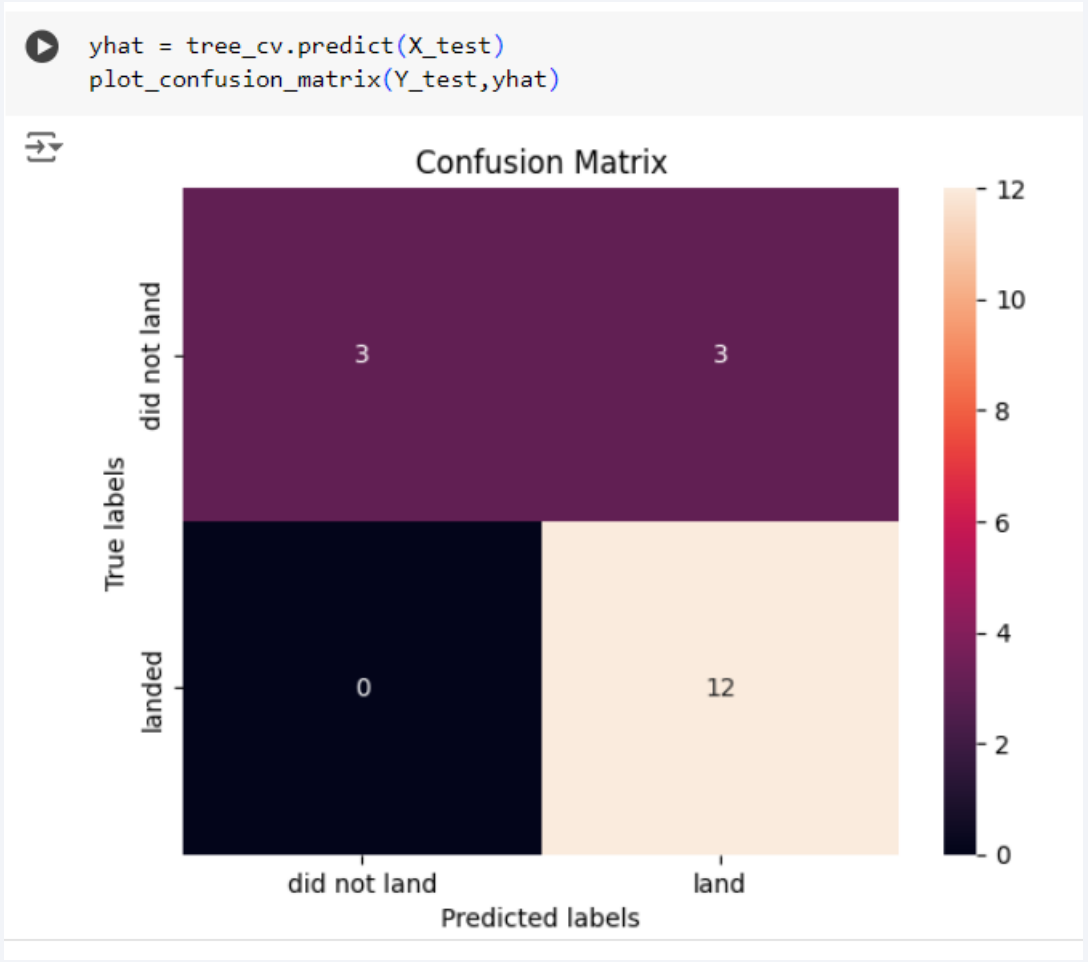
- The model with the highest accuracy is the SVM
  - It is only slightly higher than the others

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.819444	0.819444
F1_Score	0.909091	0.916031	0.900763	0.900763
Accuracy	0.866667	0.877778	0.855556	0.855556

Confusion Matrix

- This model had a high accuracy rate with a slight over confidence in the landing prediction

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- Overall Site KSC LC-39A had the best success of the launch sites
- The lighter payloads have a higher success ratio, but they also have a much higher percentage of total launches
- All sites are relatively close to the coasts while also being on the southern half of the country
- The SVM Model is the highest performing model based on accuracy

Thank you!