IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ahmed Abdelmonim Yousif Abdalla
2021/9/9

Outline

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

Executive Summary

- We collected the data from spacex rest api endpoints and web-scraped Wikipedia page that contained falcon 9 and falcon heavy data. Afterwards, we cleaned the data using pandas and numpy libraries. After processing the data, we visualized it using seaborn and folium libraries to get some insights, we also used dashboard and plotly to make interactive graphs to display the data in a more effective fashion. Ultimately, we trained four classification models which were SVM, KNN, DT and LR. Surprisingly, they all scored similar high accuracy which was 83.334%.
- With regards to the results, all launch sites needs to be located near coastlines and far away from cities for safety precautions. Moreover, they also need to be located near highways and railroads to deliver shipments and cargos to the launch location faster. In addition, launches with high payload maps tends to have a better success rate when it comes to landing the booster.

Introduction

- Launching rockets to space has always cost upward of 165 million dollars. Spacex promotes on its website that its falcon 9 rockets launches cost 62 million dollars. Most of the savings due to their capability of reusing the first stage of the falcon 9 rocket.
- In this presentation, we will predict whether the falcon 9 first stage will land successfully of not to determine the actual cost of the launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data was collected via spacex rest api and web-scaping a Wikipedia page.
- Perform data wrangling
  - Data was processed using pandas and numpy libraries.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - The data was split into train set and test
  - The perfect parameters were chosen using cross validation grid search
  - The models where trained with the best hyper-parameters

Data Collection

The data was collect via:

1- The following Spacex rest api endpoints:

A- https://api.spacexdata.com/v4/launches/past - (which contains the dataset that has the ids for the rest of endpoints)

B- https://api.spacexdata.com/v4/launchpads/(Launchpad_ids)

C- https://api.spacexdata.com/v4/rockets/(rocket_ids)

D- https://api.spacexdata.com/v4/payloads/(payload_ids)

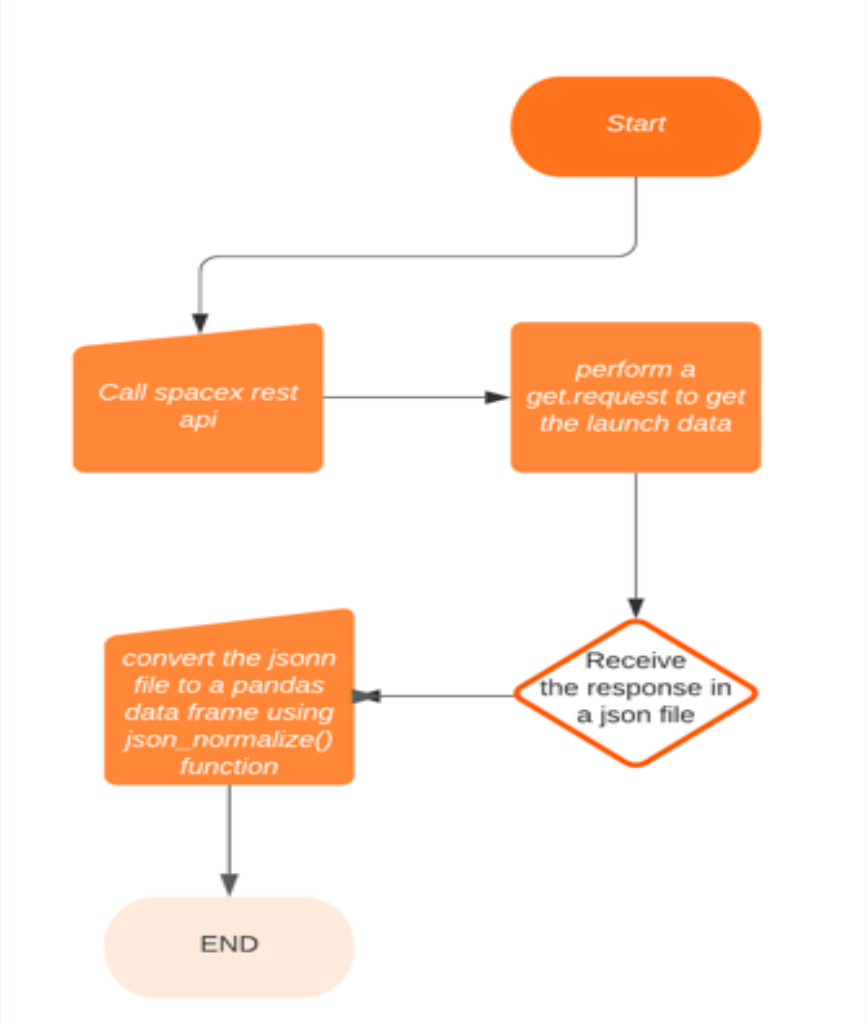
E- https://api.spacexdata.com/v4/cores/(Cores_ids)

1- Web-scraping a Wikipedia page for Falcon 9 and Falcon Heavy launches:

https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

Data Collection - SpaceX API

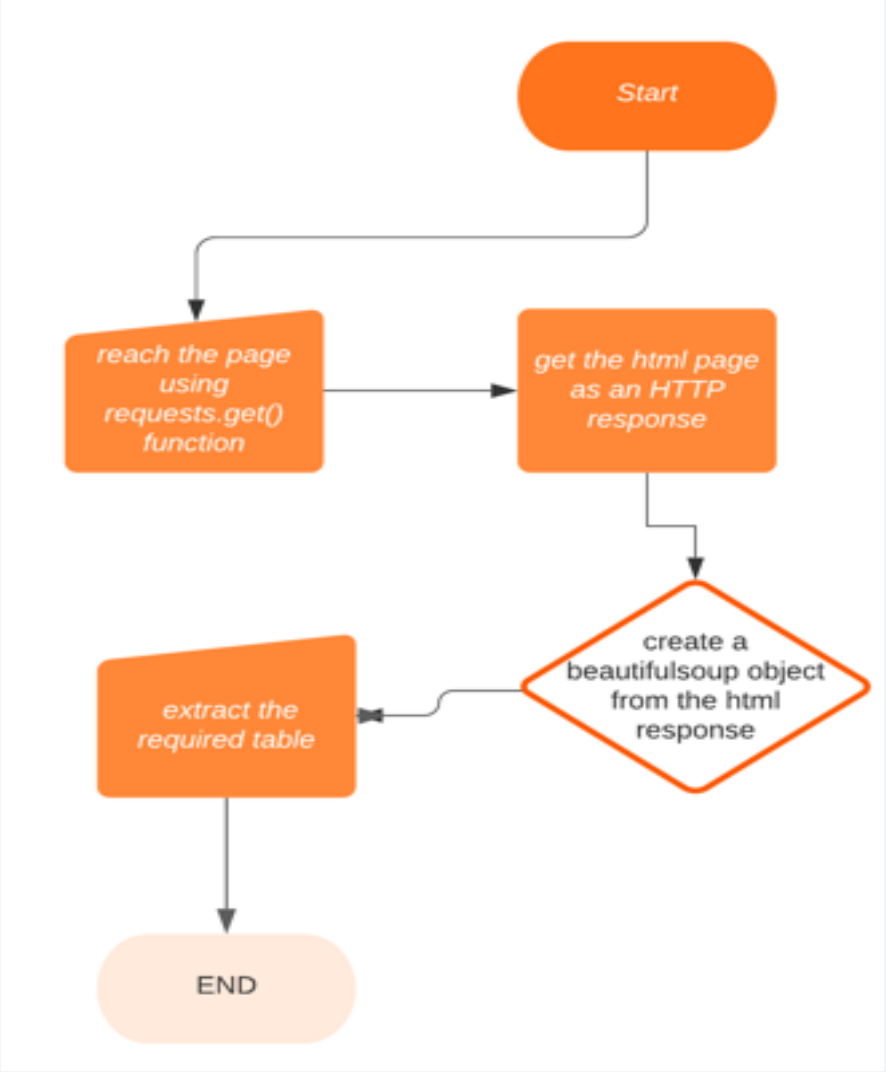
- The data was collected using spacex rest api and converted from a json file to a pandas data frame.
- The Github link for the lab is below.
- https://github.com/DS-AHMED/AHMED-ABDALLA-IBM_DataScience/blob/main/Applied%20Data%20Science%20Capstone/Data%20Collection%20API%20Lab.ipynb



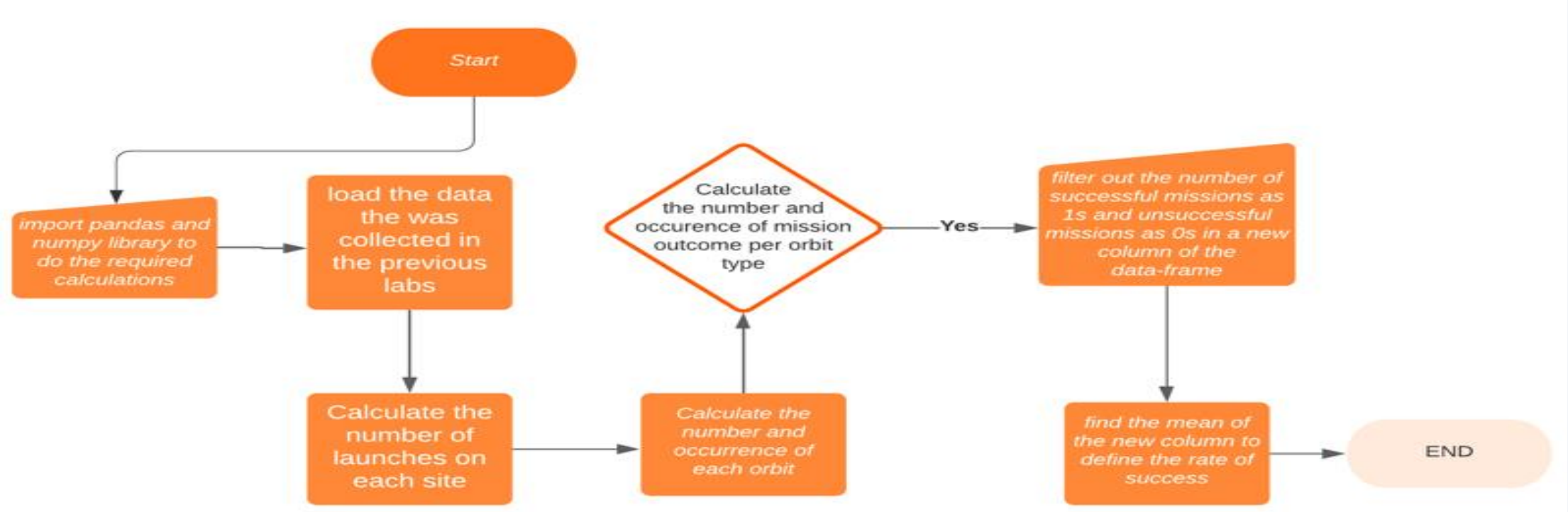
Data Collection - Scraping

- Using beautifulsoup library, we web-scraped a Wikipedia page for spacex to extract a table the contain the launches data.

- https://github.com/DS-AHMED/AHMED-ABDALLA-IBM_DataScience/blob/main/Applied%20Data%20Science%20Capstone/Data%20Collection%20with%20Web%20Scraping%20lab.ipynb



Data Wrangling



- https://github.com/DS-AHMED/AHMED-ABDALLA-IBM_DataScience/blob/main/Applied%20Data%20Science%20Capstone/EDA%20lab.ipynb

EDA with Data Visualization

- Payload Mass Vs Flight Number.
- Launch Site Vs Flight Number
- Launch Site Vs Payload Mass
- Success Rate Vs Orbit
- Orbit Vs Flight Number
- Payload Mass Vs Orbit
- Success Rate Vs Years

We compared these variables to see if there is any correlation between them regarding the success rate of the launches.

https://github.com/DS-AHMED/AHMED-ABDALLA-IBM_DataScience/blob/main/Applied%20Data%20Science%20Capstone/jupyter-labs-eda-dataviz%20%20.ipynb

EDA with SQL

https://github.com/DS-AHMED/AHMED-ABDALLA-IBM_DataScience/blob/main/Applied%20Data%20Science%20Capstone/EDA%20with%20SQL%20lab.ipynb

- %sql select * from spacextbl1
- %sql select UNIQUE(launch_site) from SPACEXTBL1;
- %sql select * from spacextbl1 where launch_site like '%CCA%' limit 5
- %sql select sum(payload_mass__kg_) as NASA_total_payload_mass from spacextbl1 where customer = 'NASA (CRS)‘
- %sql select avg(payload_mass__kg_) as avg_payload_mass from spacextbl1 where booster_version = 'F9 v1.1‘
- %sql select min(date) from spacextbl1 where landing__outcome like '%pad%‘
- %sql select booster_version from spacextbl1 where landing__outcome = 'Success (drone ship)' and 4000<payload_mass__kg_<6000 ;
- %sql select count(mission_outcome) as total_outcome_number from spacextbl1;
- %sql select booster_version from spacextbl1 where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl1);
- %sql select DATE , landing__outcome, booster_version, launch_site from spacextbl1 where landing__outcome = 'Failure (drone ship)' and DATE like '%2015%' ;
- %sql select landing__outcome, \ count(landing__outcome) as "no of Lanidng outcome" from spacextbl1 \ where DATE between '2010-06-04' and '2017-03-20'\ group by landing__outcome\ order by count(landing__outcome) desc ;

Build an Interactive Map with Folium

- We use folium.Circle object to add a highlighted circle area on a specific coordinate.
- We used folium.map.Marker object to add a text label on specific coordinates.
- We used a MarkerCluster() object to mark the successful launches with a green color and the unsuccessful ones with a red color.
- We added a MousePosition() object to locate the coordinates on the map automatically.
- We added a folium.PolyLine() object to draw lines to measure related distances
- Explain why you added those objects
- https://github.com/DS-AHMED/AHMED-ABDALLA-IBM_DataScience/blob/main/Applied%20Data%20Science%20Capstone/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

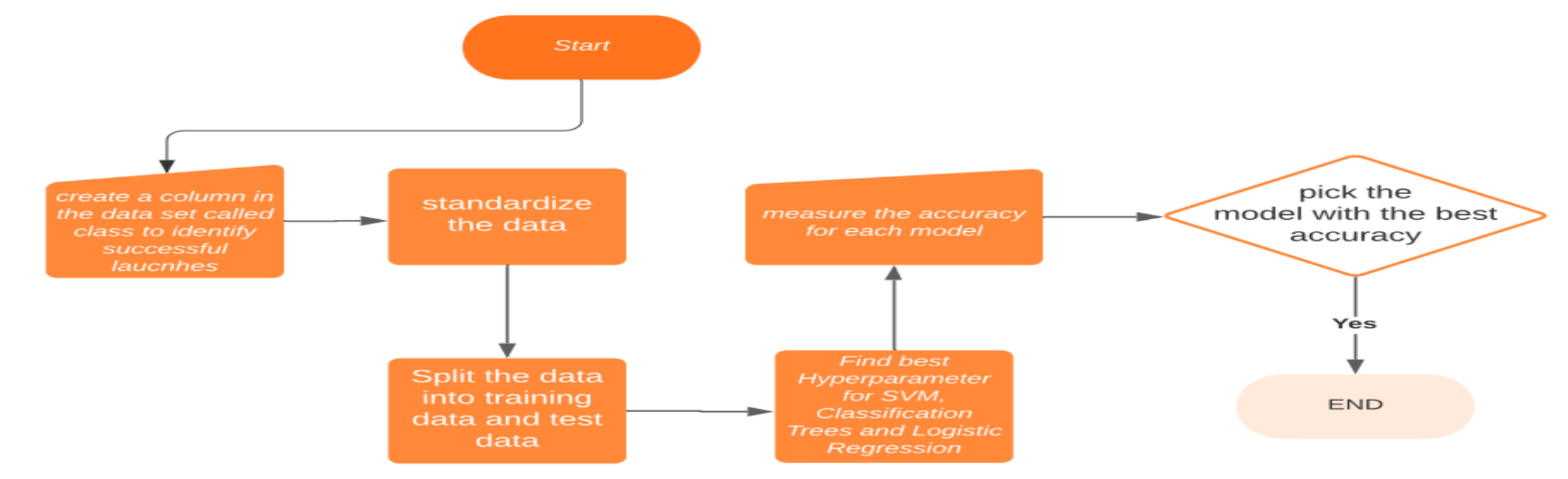
A- A pie-chart to display the success rate for all sites or each site individually to find which site has the best success rate, the site names are:

1- CCAFS LC-40. 2- CCAFS SLC-40. 3- KSC LC-39A. 4- VAFB SLC-4E.

B- A scatter map to see the effect of payload mass on the success rate of the mission.

- https://github.com/DS-AHMED/AHMED-ABDALLA-IBM_DataScience/tree/main/Applied%20Data%20Science%20Capstone/dashboard

Predictive Analysis (Classification)



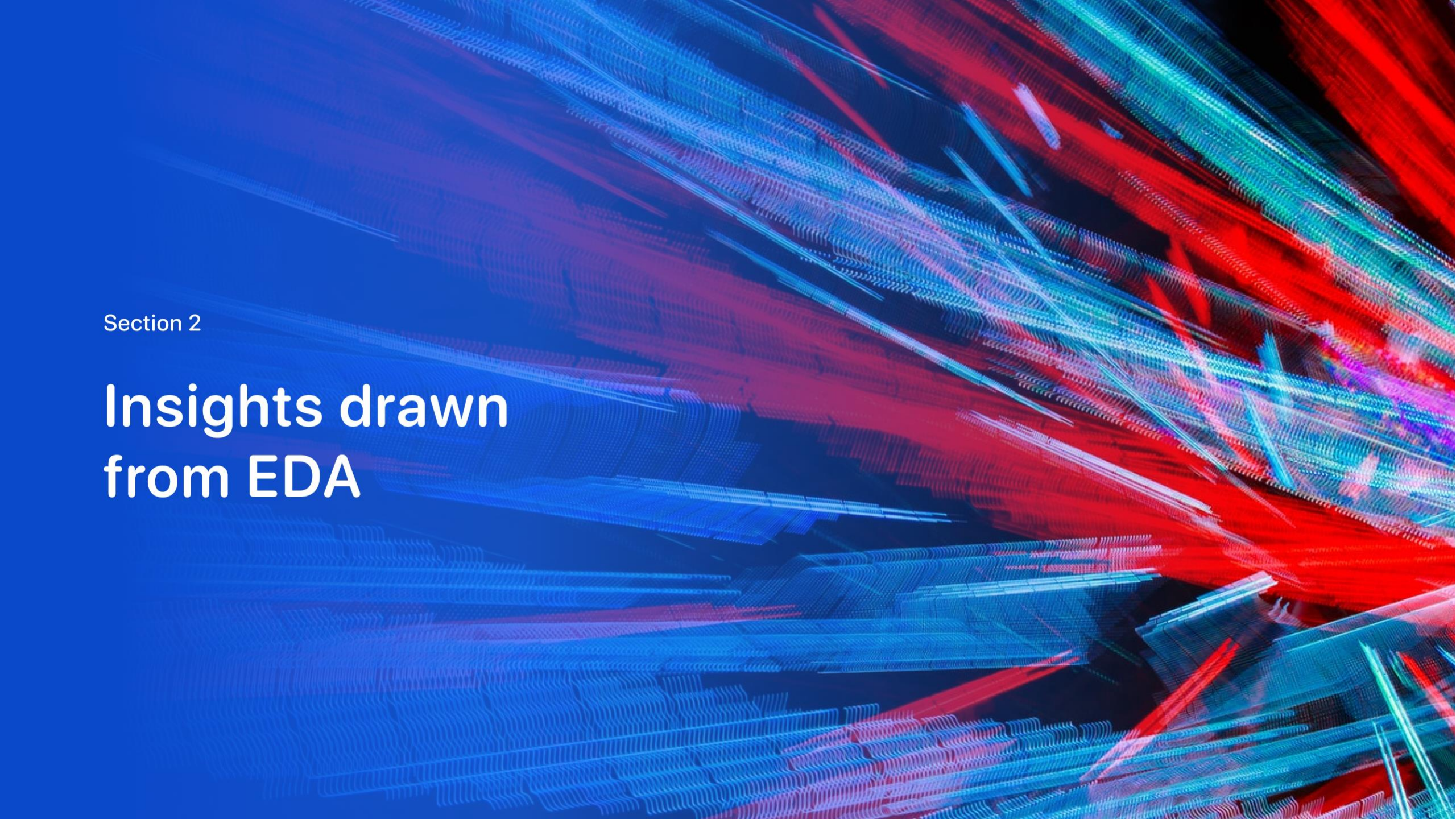
After defining successful and unsuccessful launches with 1s and 0s, we standardized the variables to makes their weight when training the model similar, then split the data to do A-B testing to define the best accuracy, then before training each model we used **gridsearchcv** to find the best hyper-parameters for each model. For testing the accuracy, the sued jaccard and f1-score, we found that other than Decision tree model, support vector machine, K nearest neighbor and logistic regression models all performed similarly well. In conclusion, there're good models to predictive analysis except for Decision tree model.

Results

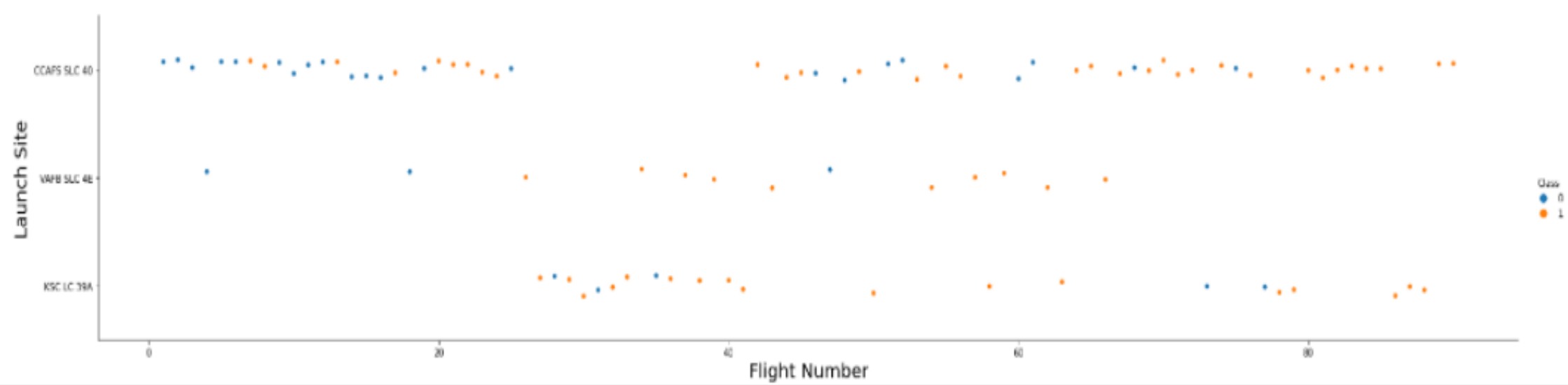
- The more launches spacex performed the better their success rate was.
- The more payload mass the rockets carried the better the success rate was.
- The less launches had been performed the better the success rate for each orbit was, except for flights after flight number 60, all orbits in general had a high success rate.
- Since 2013 till 2020, spacex witnessed a steady increase in launches success rate.
- Launch locations should be near highways and railroads to move cargos faster, they also should be located far-away from cities and near coastlines to avoid casualties when launches fail.
- VAFB SLC-4E location has the highest success rate which was 76.9%
- SVM, KNN and LR all scored 81% in jaccard accuracy and 80% in F1-score which makes them perfect models to predict future flights success or failure

Section 2

Insights drawn from EDA

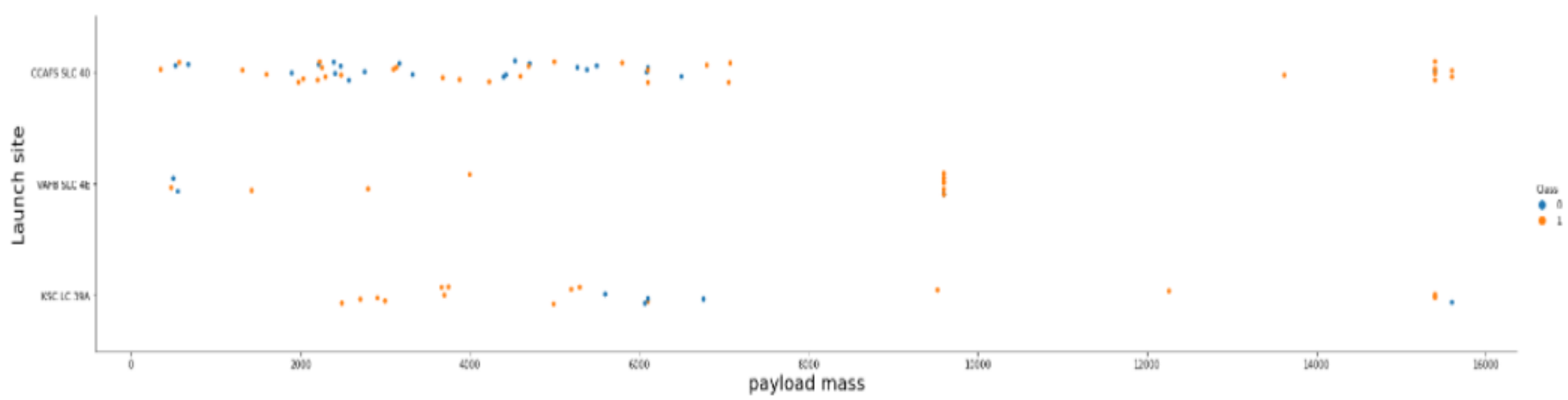


Flight Number vs. Launch Site



- As flight number increases, the better the success rate for launches. CCAFS SLC-40 had the highest number of flights, it started with a very low success rate but then gradually started getting a better success rate. KSC LC-39A and VAFB SLC-4E had almost similar success rate, yet the latter has many fewer flights than the former.

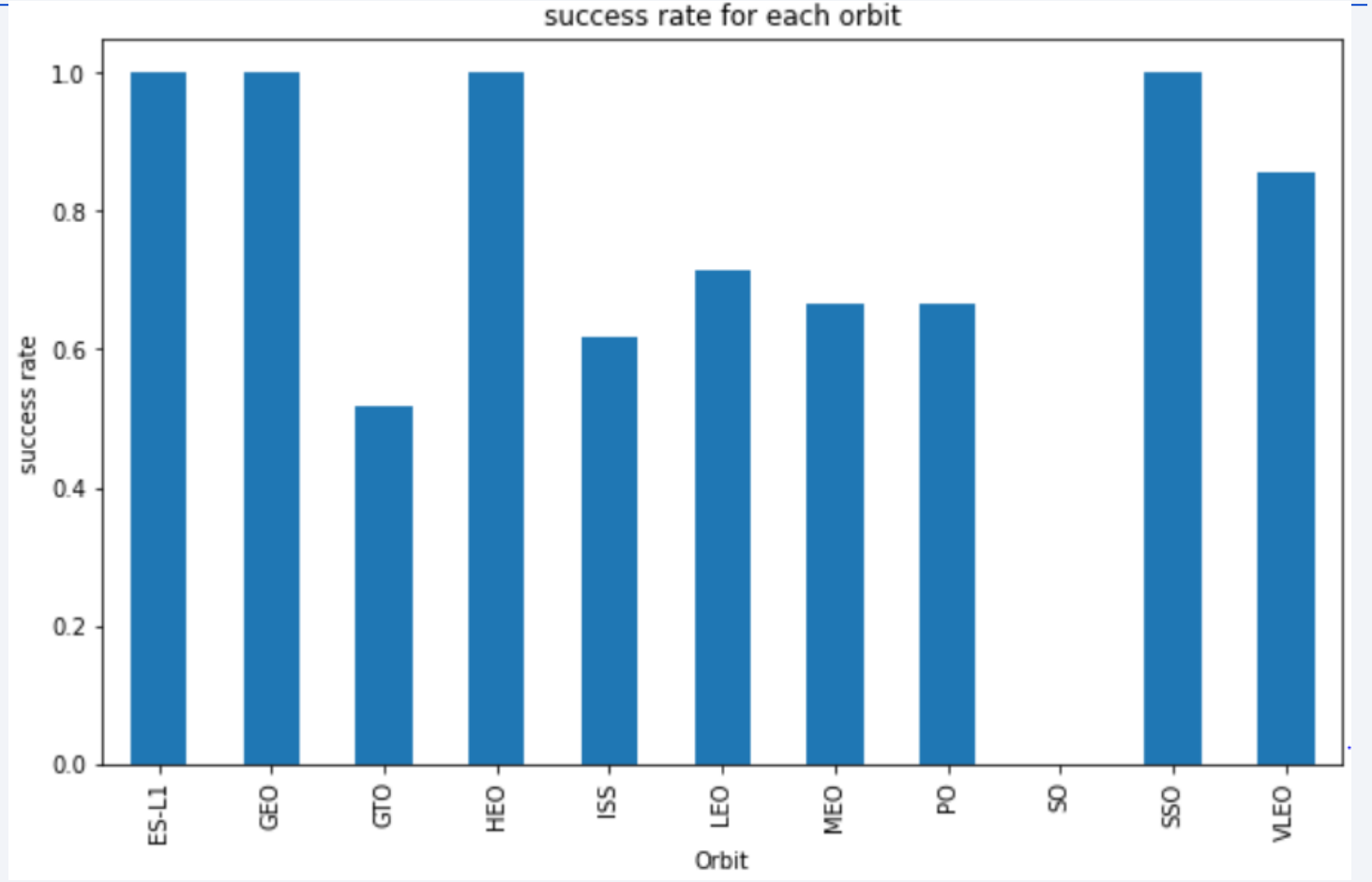
Payload vs. Launch Site



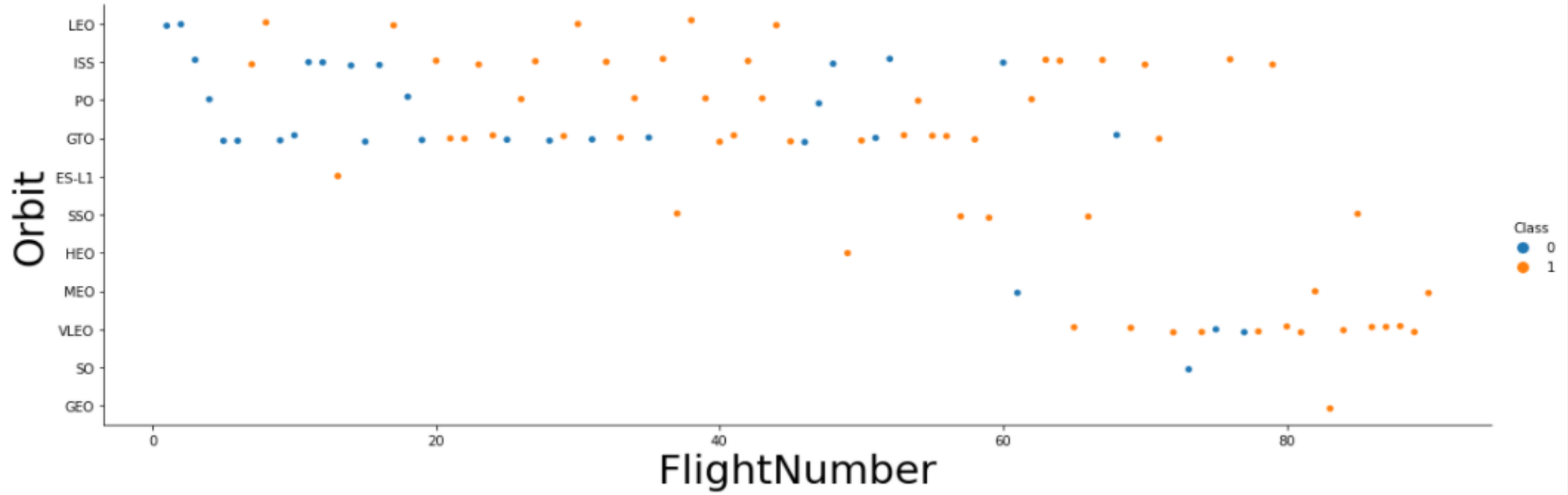
- It’s quite obvious that the higher the payload mass the better the success rate for launches regardless of which location the rocket it launched from.

Success Rate vs. Orbit Type

- Regardless of the number of occurrence in each orbit, ES-L1, GEO, HEO and SSO had a success rate of 100%. VLEO had success rate that is just above 80% and the rest of orbits had a success rate between 50% to 70%.

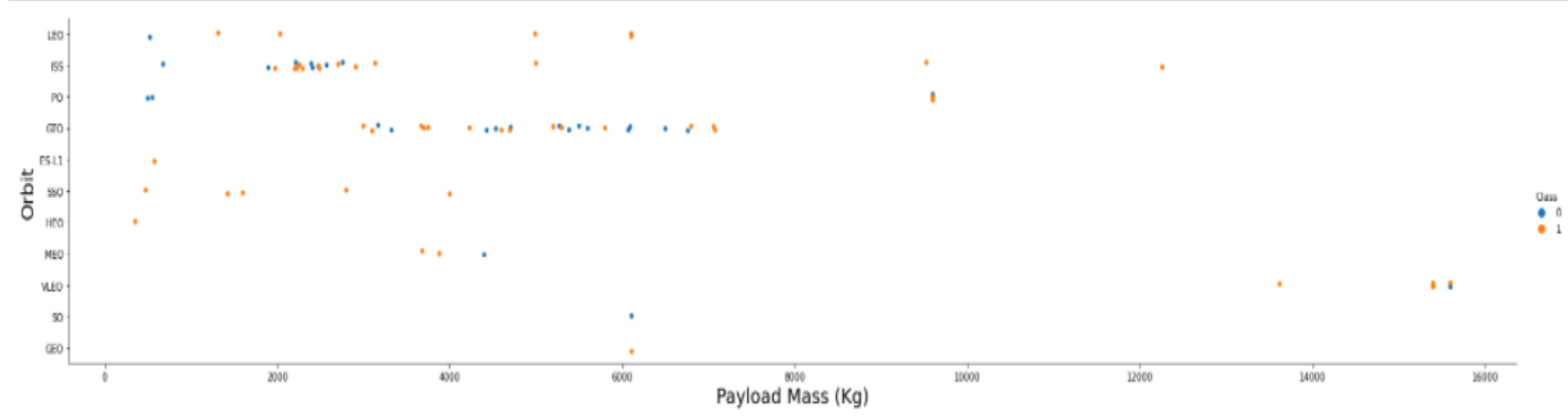


Flight Number vs. Orbit Type



- Overall, orbits don't have a strong correlation with flight number when it comes to the success rate. However, the launches on the first 4 orbits seems to represent the majority of the launches which indicates a weak positive correlation between the flight numbers and orbits.

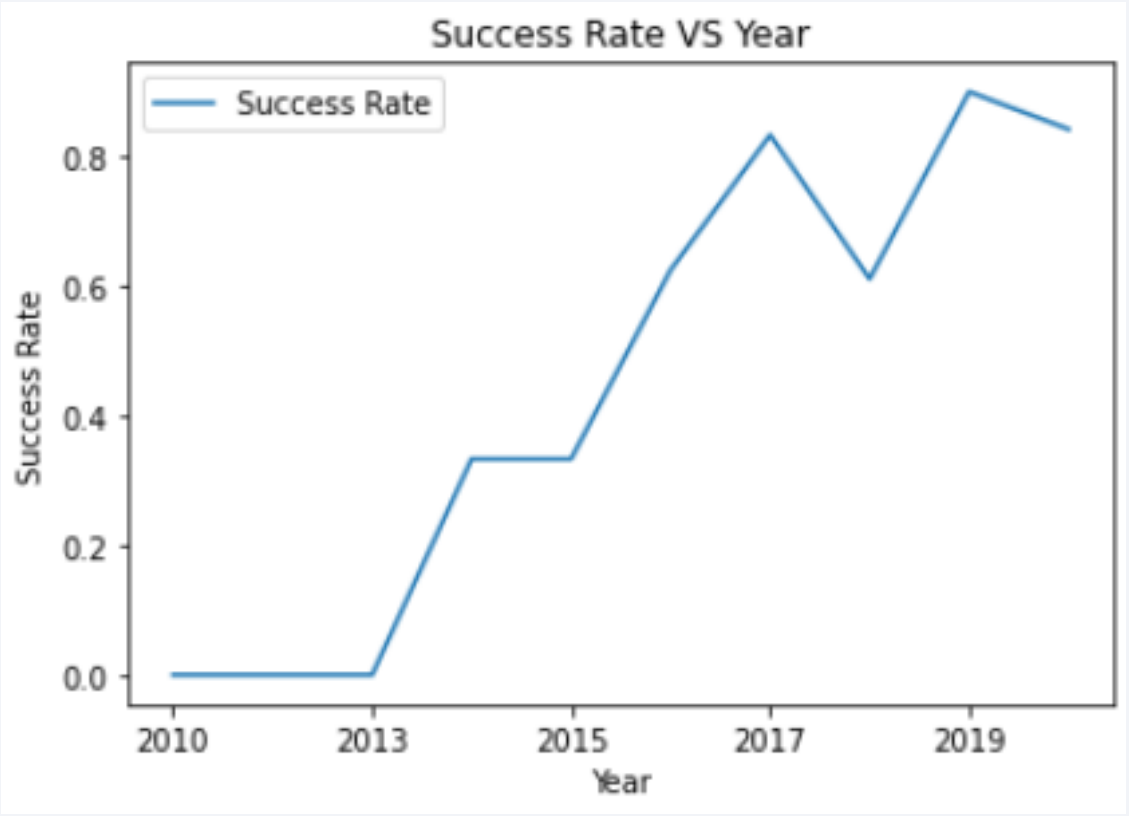
Payload vs. Orbit Type



- There is now correlation between payload mass and any orbit when it comes to the success rate.

Launch Success Yearly Trend

- It is readily apparent that since 2013 success rate is almost linearly increasing which means, the more years of experience the company has, the higher success rate it can score regarding the landing of the first stage.



All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

In [9]:
```
%sql select UNIQUE(launch_site) from SPACEXTBL1;
```

```
 * ibm_db_sa://fdh07761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[9]:

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

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

In [10]:
```
%sql select * from spacextbl1 where launch_site like '%CCA%' limit 5
```

```
 * ibm_db_sa://fdh07761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[10]:

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

Display the total payload mass carried by boosters launched by NASA (CRS)

In [11]:
```
%sql select sum(payload_mass__kg_) as NASA_total_payload_mass    from spacextbl1 where customer = 'NASA (CRS)'
```

```
 * ibm_db_sa://fdh07761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[11]:

nasa_total_payload_mass
45596

Average Payload Mass by F9 v1.1

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

In [12]:
```
%sql select avg(payload_mass__kg_) as avg_payload_mass from spacextbl1 where booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://fdh07761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[12]:

avg_payload_mass
2928

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

In [30]:
```
%sql select min(date) from spacextbl1 where landing__outcome like '%pad%'
```

```
 * ibm_db_sa://fdh07761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[30]:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [43]:
```
%sql select booster_version from spacextbl1 where landing__outcome = 'Success (drone ship)' and 4000<payload_mass__kg_<6000 ;
```

```
 * ibm_db_sa://fdh07761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[43]:

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

In [46]:
```
%sql select count(mission_outcome) as total_outcome_number from spacextbl1;
```

```
 * ibm_db_sa://fdh07761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[46]:

total_outcome_number
101

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [54]:
```
%sql select booster_version from spacextbl1 where  payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl1);
```

```
 * ibm_db_sa://fdh07761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[54]:

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

Task 9

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for the in year 2015

In [61]:
```
%sql select DATE , landing__outcome, booster_version, launch_site from spacextbl1 where landing__outcome = 'Failure (drone ship)'
and DATE like '%2015%'  ;
```

```
 * ibm_db_sa://fdh07761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

Out[61]:

DATE	landing__outcome	booster_version	launch_site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

Rank the 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

In [92]:
```
%sql select landing__outcome, \
count(landing__outcome) as "no of Lanidng outcome" from spacextbl1 \
where DATE between '2010-06-04' and '2017-03-20'\
group by landing__outcome\
order by count(landing__outcome) desc  ;
```

```
 * ibm_db_sa://fdh07761:***@54a2f15b-5c0f-46df-8954-7e38e612c2bd.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32733/bludb
Done.
```

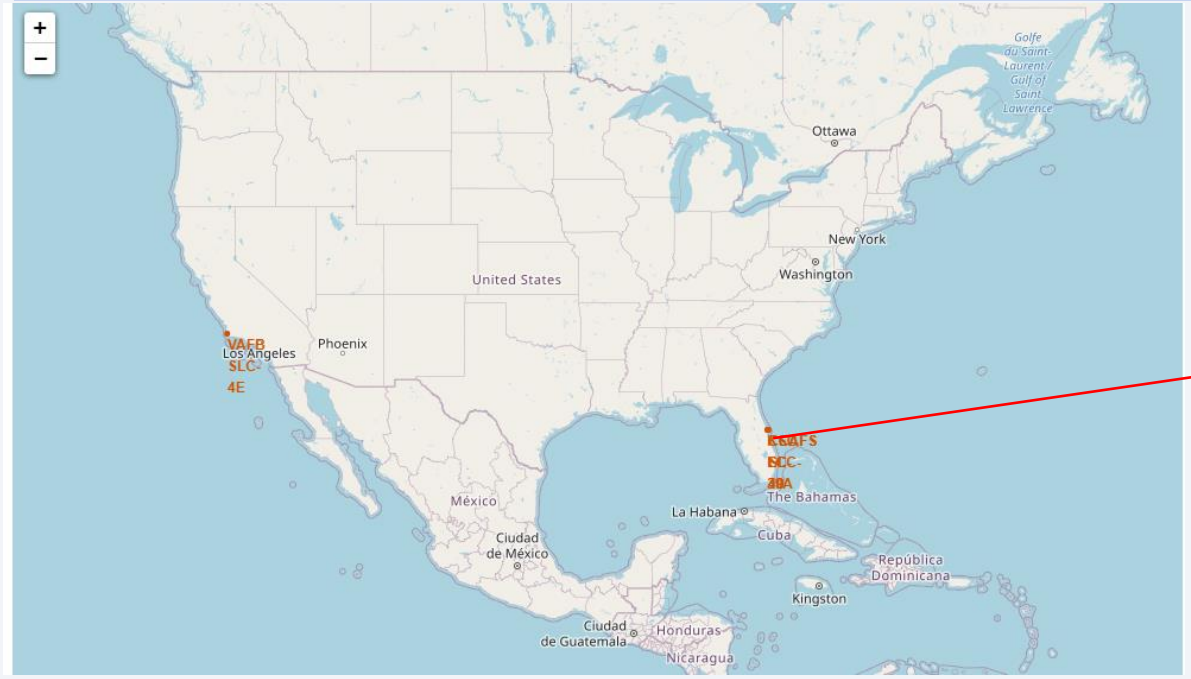
Out[92]:

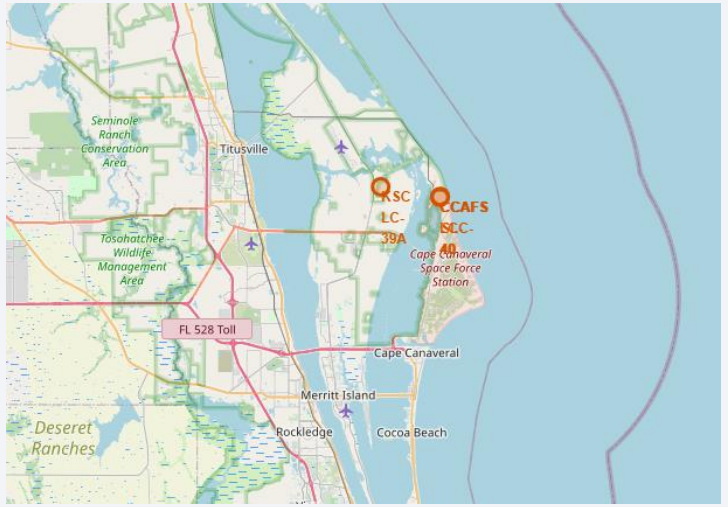
landing__outcome	no of Lanidng outcome
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

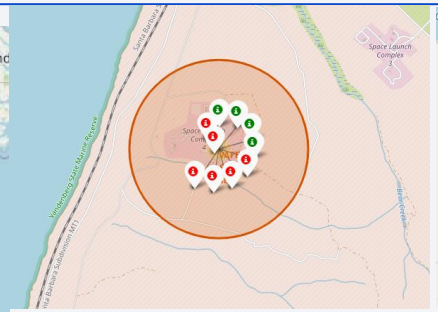
Launch Sites Location



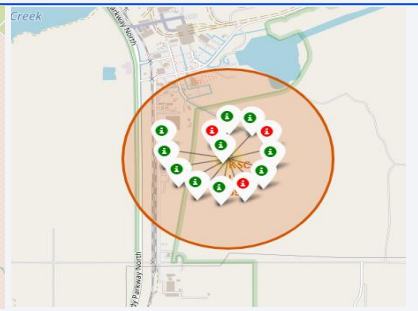


There are no launch sites in midland, this is to avoid casualties in case in of failed launches. Moreover, three out of the four launch sites are in one location, which means it's the optimum location for launching rockets for spacex.

Success rate for each launch location

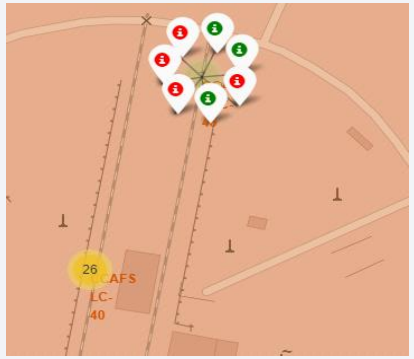


VAFB SLC-4E



KSC LC-39A

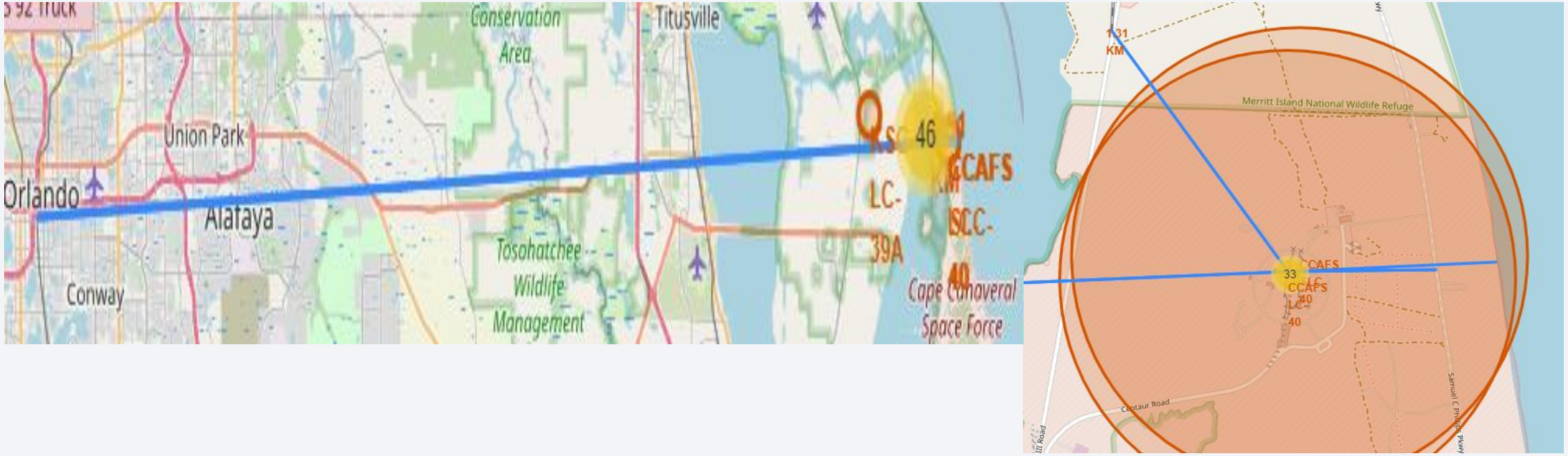
CCAFS LC-40



CCAFS SLC-40

- CCAFS LC-40 had the highest number of launches and just bellow 30% success rate.
- CCAFS SLC-40 had the lowest number of launches with approximately a success rate of 40%.
- VAFB SLC-4E and KSC LC-39A had approximately a similar number of launches with a success rate of 40% and 77% respectively.

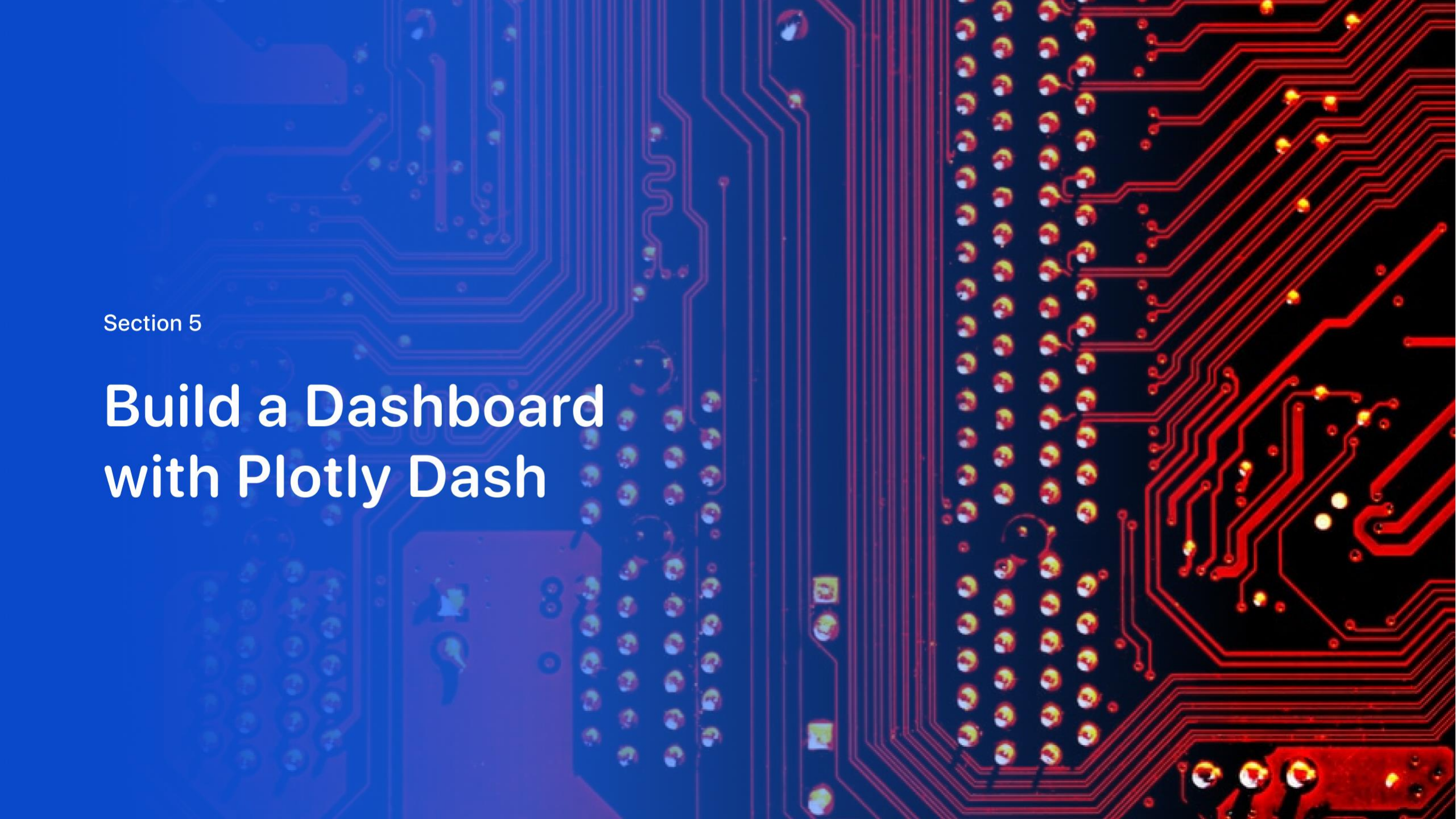
launch site to its proximities



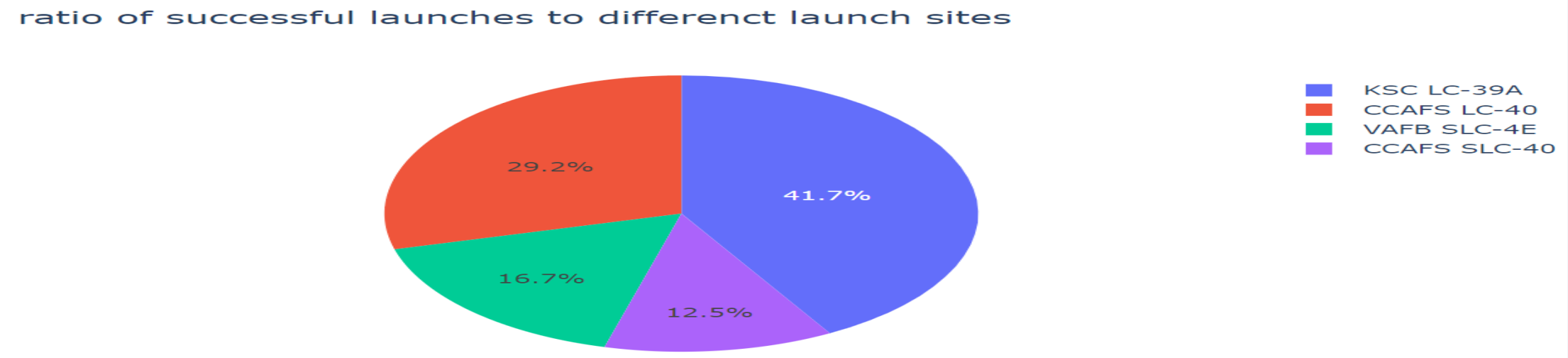
- For all locations, launch sites are located with close proximity to highways and railroads to deliver cargos faster and near coastlines and far away from cities for safety precautions.
- Closest city is ≈ 78.68 Km away, closest coastline is ≈ 0.91 Km away, closest highway is ≈ 0.65 Km away, closest railroad is ≈ 1.31 away from CCAFS LC-40.

Section 5

Build a Dashboard with Plotly Dash

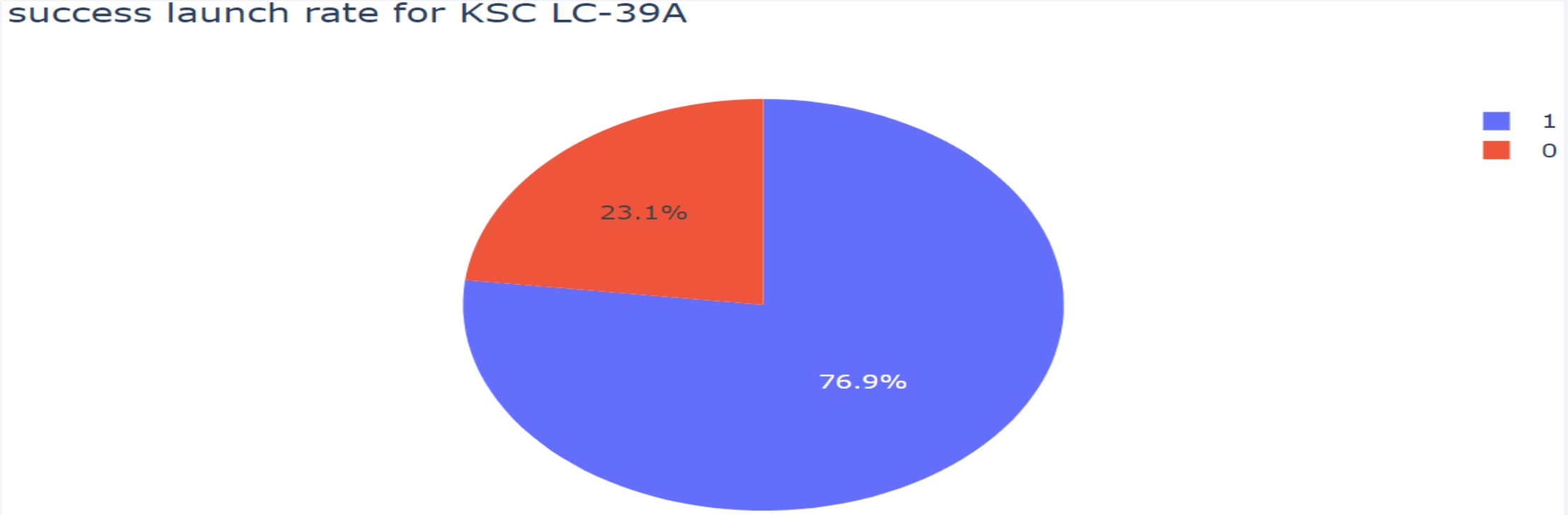


Launch success count for all sites



- This pie-chart displays the launch success ratio in comparison among all launch sites. KSC LC-39A had the highest success ratio which was 41.7%.

Launch success for KSC LC-39A



- This pie-chart shows the launch success rate of KSC LC-39A launch site, which scored the highest success rate among all other launch sites.

Payload vs. Launch Outcome scatter plot for all sites

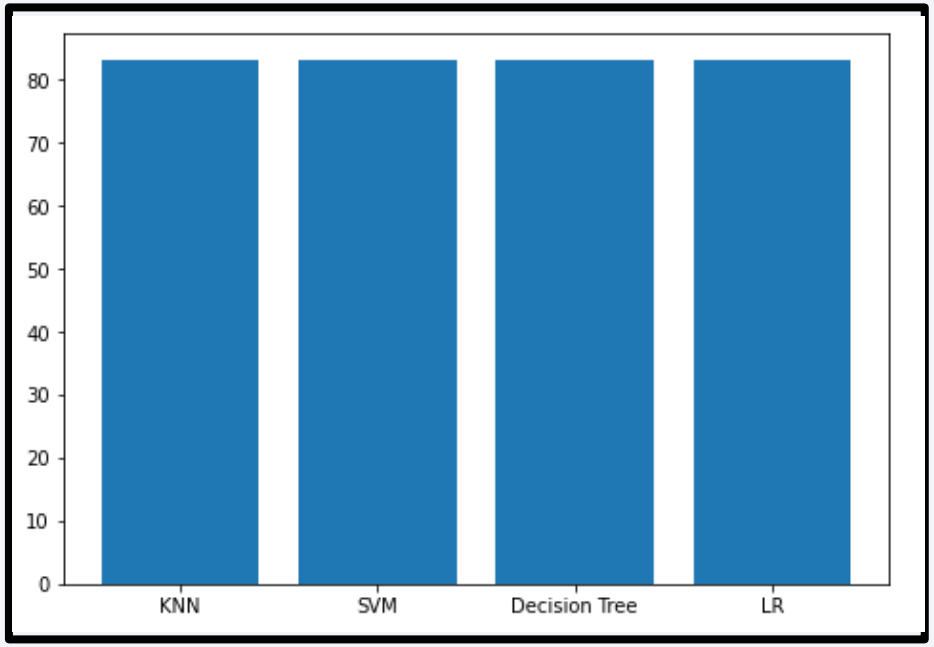
This interactive dash-board scatter plot explain the relationship between payload mass and launch success outcome. It readily shows that the heavier the payload mass, the higher the possibility that the launch will succeed.

Section 6

Predictive Analysis (Classification)

Classification Accuracy

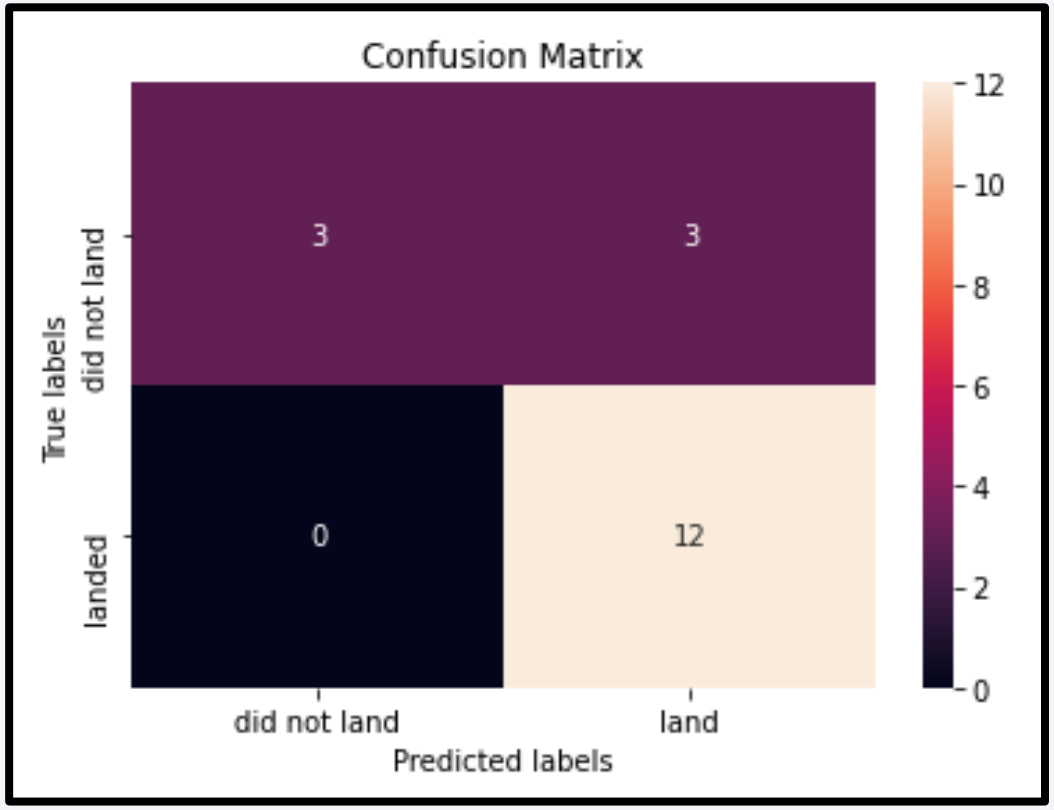
	Predictive Model	Score
0	KNN	8.3334
1	SVM	8.3334
2	Decision Tree	8.3334
3	LR	8.3334



- We split the dataset (which contain 83 features) into a training set with 80% of the data and test set with 20% of the data to train and test the models .
- Luckily, all the model performed similarly well. They all scored an accuracy of 83.334%.
- To get this accuracy, we performed a grid search to find the best hyper-parameters.

Confusion Matrix

- This accuracy matrix displays the accuracy of the model with respect to the test data.
- Top lest box shows the launches that actually didn’t land and was predicted correctly by the model.
- Bottom left box shows the launches that actually landed and was predicted correctly by the model.
- The other diagonal shoes vice versa.



Conclusions

According to data, to ensure the success of the launches, a few things need to be done:

1- increase the payload mass when launching the rockets.

2- launch sites needs to be located far away from cities and close to coastlines for safety precautions.

3- launch sites also needs to be located near highways and railroads to move cargos and equipment faster.

Appendix

- Spacex rest api end point.
- List of Falcon 9 and Falcon Heavy launches on Wikipedia.
- Processed data used to train the classification models.

Thank you!