Winning Space Race with Data Science

Alexandre Michel Maul
March/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Access to Space X API and data collection (get request)
  - Data collection and data cleansing
  - Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  Space X advertises Falcon 9 rocket launches with a cost of 62 million dollars. Other providers cost upward of 165 million dollars each. Much of the savings is because Space X can reuse the first stage rocket.

  Therefore, if we can determine if the first stage will land successfully, we can determine the potential profit of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

  The next slides present a machine learning pipeline to predict if the first stage will land successfully.

- Answers that we can provide:
  - Which factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - Which operating conditions are likely to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping from Wikipedia.
- Perform data wrangling
  - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data were analyzed and models tested to identify patterns and the key factors for successful rocket landing

Data Collection

- Data were collected from two main sources
  - Space X API
  - Data from Space X API was collected using a get request.
    - Data collected (.json file) was decoded using .json() function call and converted into a dataframe.
  - Data was scanned, checked for missing values and cleaned, filling in missing values where necessary.
  - Wikipedia website
  - Wikipedia was scrapped with BeautifulSoup.
    - Data collected/extracted (HTML table) was parsed converted to a dataframe.

Data Collection – SpaceX API

- We used the get request to the SpaceX API to collect data, clean the requested data and did some basic data wrangling and formatting.

```
In [7]: spacex_url="https://api.spacexdata.com/v4/launches/past"

In [8]: response = requests.get(spacex_url)

        Check the content of the response

In [9]: print(response.content)

        b'[{"fairings":{"reused":false,"recovery_attempt":false,"recovered":false,"ships":[]},"li
        o.png","large":"https://images2.imgbox.com/40/e3/GypSkayF_o.png"},"reddit":{"campaign":nu
```

```
In [10]: static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-S

         We should see that the request was successfull with the 200 status response code

In [11]: response.status_code

Out[11]: 200

         Now we decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_norma

In [12]: # Use json_normalize meethod to convert the json result into a dataframe
         data = pd.json_normalize(response.json())
```

```
In [41]: data_falcon9.isnull().sum()

Out[41]: FlightNumber      0
         Date              0
         BoosterVersion    0
         PayloadMass       0
         Orbit             0
         LaunchSite        0
         Outcome           0
         Flights           0
         GridFins          0
         Reused            0
         Legs              0
         LandingPad       26
         Block             0
         ReusedCount       0
         Serial            0
         Longitude         0
         Latitude          0
         dtype: int64
```

- https://github.com/Ammaul/SpaceX/blob/master/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

- We did a web scrapping to get Falcon 9 launch records with BeautifulSoup from wikipedia

In [6]:
```
# use requests.get() method with the provided static_url
# assign the response to a object

response = requests.get(static_url).text
```

Create a `BeautifulSoup` object from the HTML `response`

In [7]:
```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response, 'html.parser')
```

In [8]:
```
# Use soup.title attribute
print(soup.title)
```

```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

In [51]:
```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each value t
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]

launch_dict
```

In [41]:
```
df.head(5)
```

Out[41]:

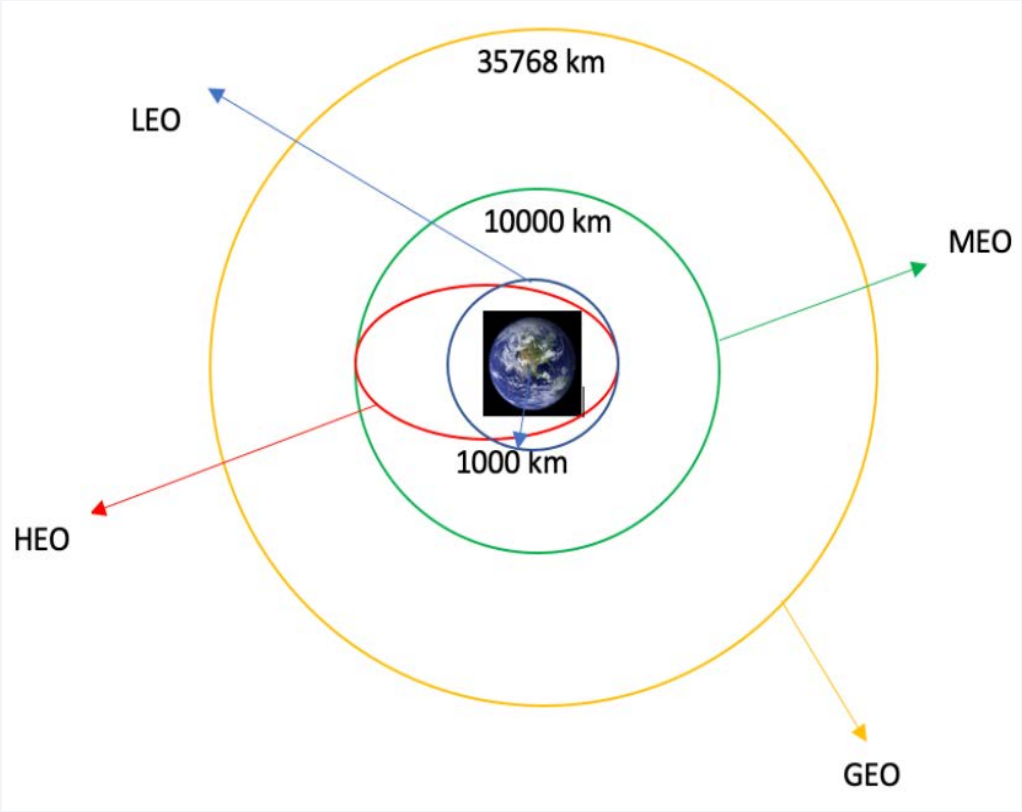
	Flight No.	Launch site	Payload	Payload mass	Orbit
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO
1	2	CCAFS	Dragon	0	LEO
2	3	CCAFS	Dragon	525 kg	LEO
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO

- https://github.com/Ammaul/SpaceX/blob/master/jupyter-labs-webscraping.ipynb

Data Wrangling

- We did some exploratory data calculations and analysis.
- As a first approach were evaluated number of launches for each site, number and occurrence of each orbits
- To smooth further analysis, some labels were defined and created to rank landing outcomes.

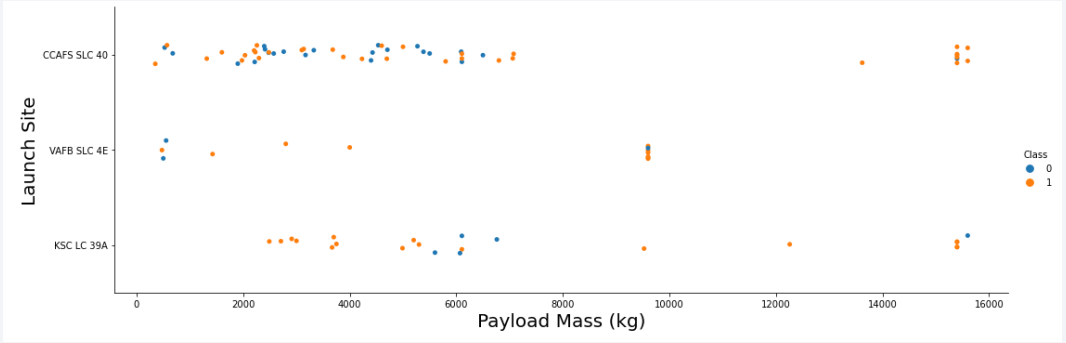
```
In [28]: for i,outcome in enumerate(landing_outcomes.keys()):
             print(i,outcome)

0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 False Ocean
6 None ASDS
7 False RTLS
```

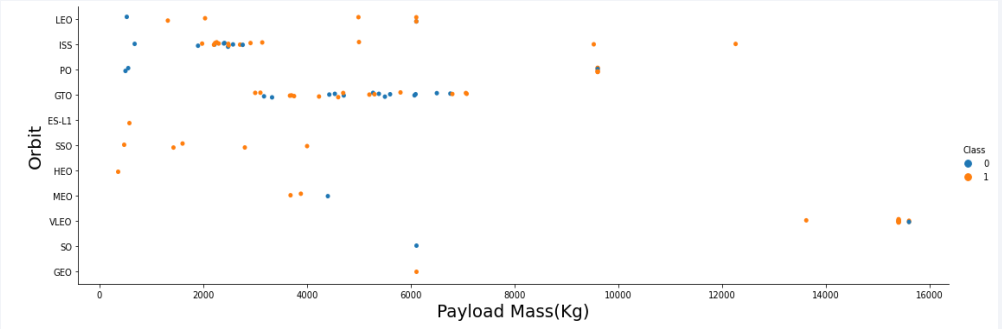


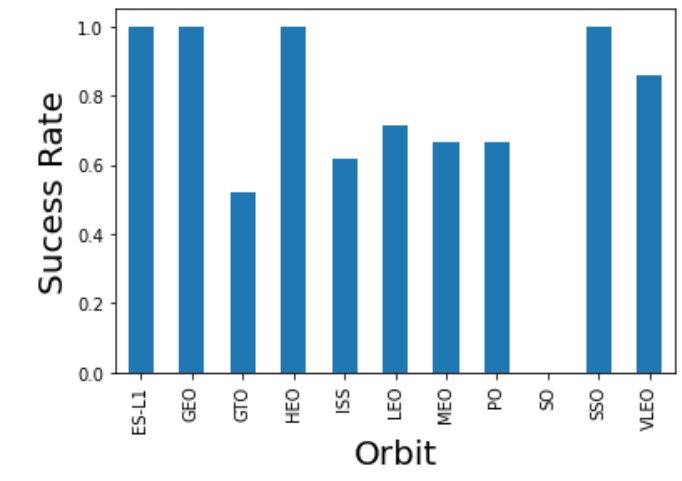
- https://github.com/Ammaul/SpaceX/blob/master/labs-jupyter-spacex-Data%20wrangling.ipynb

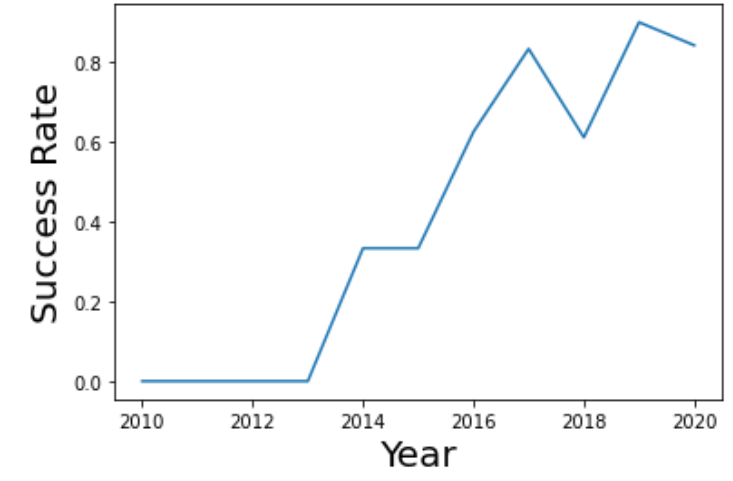
EDA with Data Visualization

- Some variables were plotted and compared to allow a better understanding between key variables: relationship between flight number and launch site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend, among others.









- https://github.com/Ammaul/SpaceX/blob/master/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- We loaded the SpaceX dataset into a DB2 database and queried data using the jupyter notebook.
- The queries provided some insight from the data, mainly:
  - Unique launch sites names.
  - Payload launched by NASA (CRS)
  - Payload for booster version F9 v1.1
  - Number of successful and failure mission outcomes
  - Number of failed landing outcomes in drone ship, their booster version and launch site names.

Out[29]:

payloadmass
45596

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

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

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

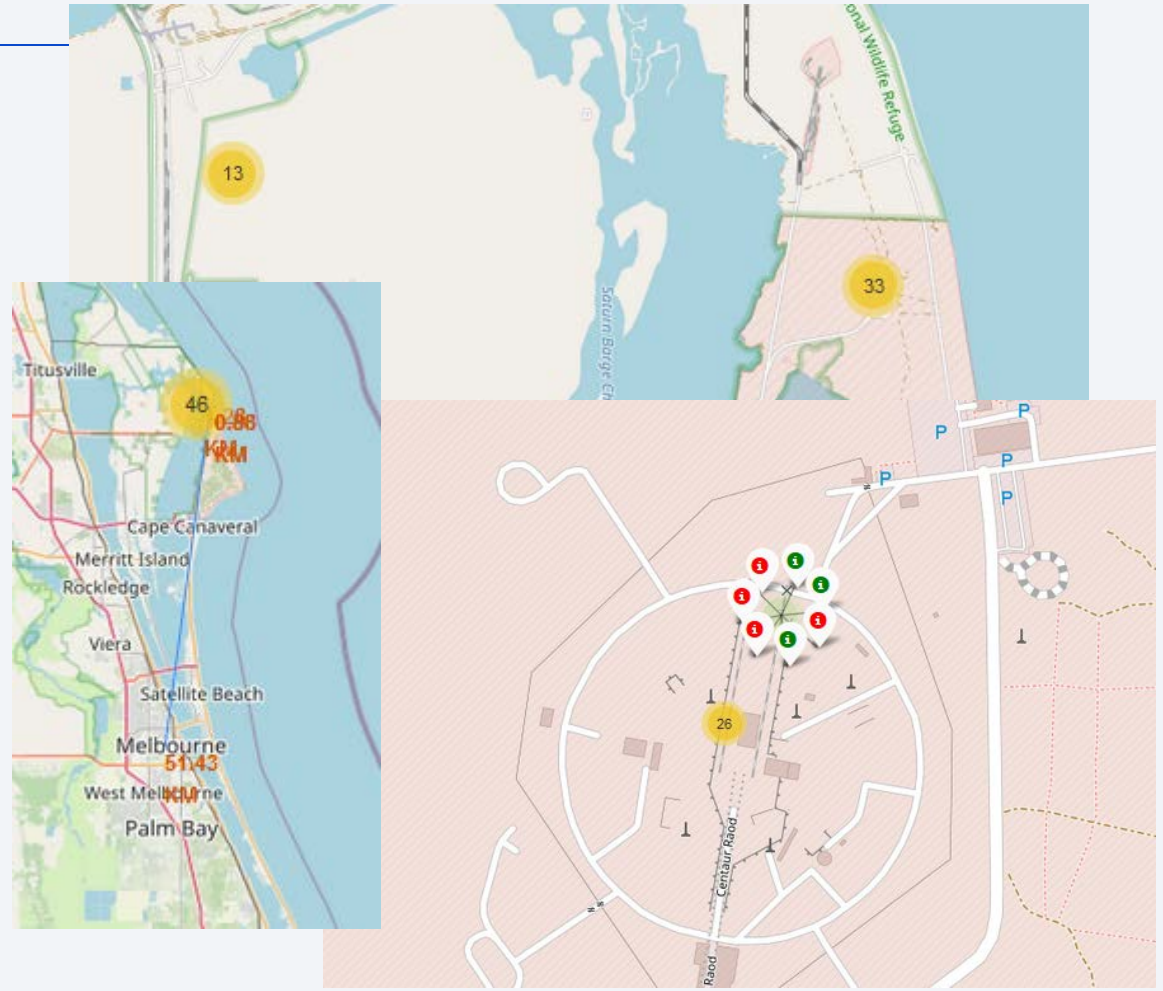
booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- https://github.com/Ammaul/SpaceX/blob/master/jupyter-labs-eda-sql-coursera.ipynb

Interactive Map with Folium

- Launch sites marked with objects such as markers, circles, lines were built
  - Success or failure launches were indicated for each site using folium map.
- Launch outcomes (failure or success) were sorted to 0 (failure) and 1 (success).
  - Marker clusters were included at launch sites.
- Distances between a launch site and proximities were calculated:
  - Nearby citties, railways, highways and coastlines.



- https://github.com/Ammaul/SpaceX/blob/master/lab_jupyter_launch_site_location.ipynb

Plotly Dash Dashboard

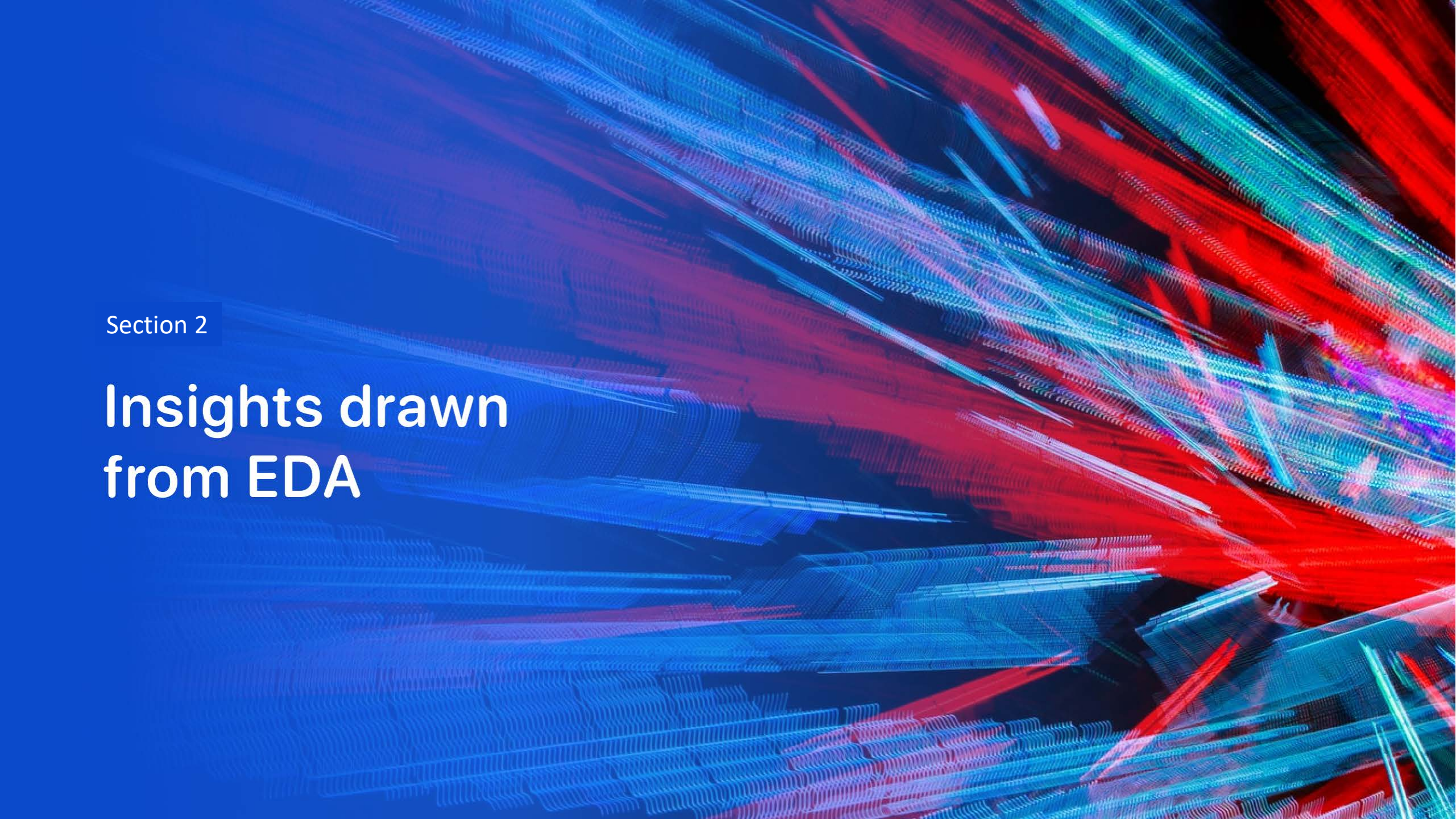
- An interactive dashboard with Plotly dash was built using a dedicated Cloud IDE.
- Data was loaded using “wget” command line from a dataset spacex_launch_dash.csv.
- Two diferente data analysis were performed:
- a) A dropdown menu was created to render a pie chart showing successful rate for each launch site.
- B) A slider bar was used to select a payload range. Based on this selection, a scatterplot shows the mission outcome for each booster version used.

- https://github.com/Ammaul/SpaceX/blob/master/Plotly%20Dash%20Dashboard

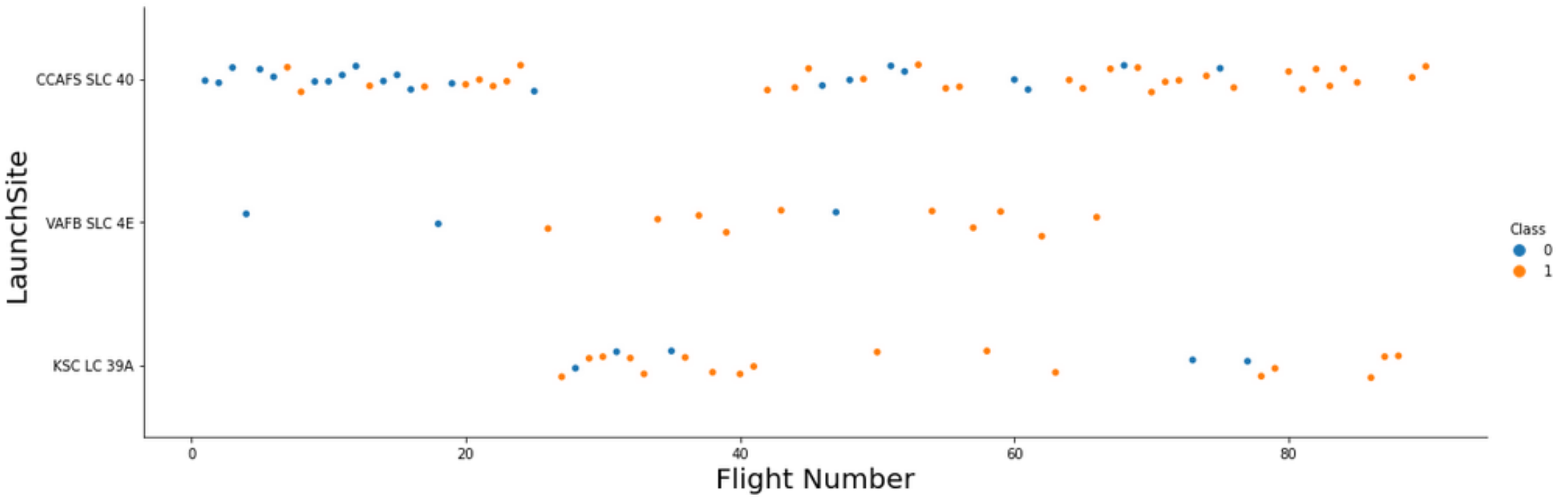
Predictive Analysis (Classification)

- Data was loaded using read function. After that, they were analyzed using numpy and pandas, transformed and split into training and testing clusters.
- Three different machine learning models were fitted. The best tune parameters for each model were selected using GridSearchCV.
- The confusion matrix was plotted for all models, with equal results for all models.
- Finally, accuracy was used to select the best model: Decision tree model.

- https://github.com/Ammaul/SpaceX/blob/master/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

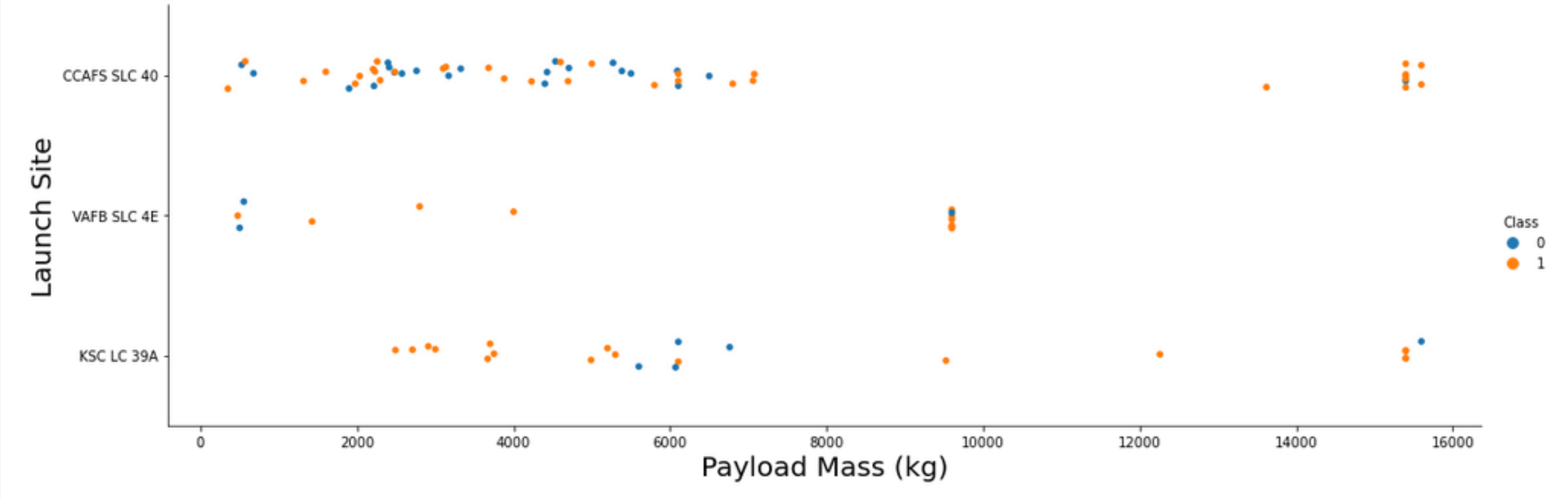

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



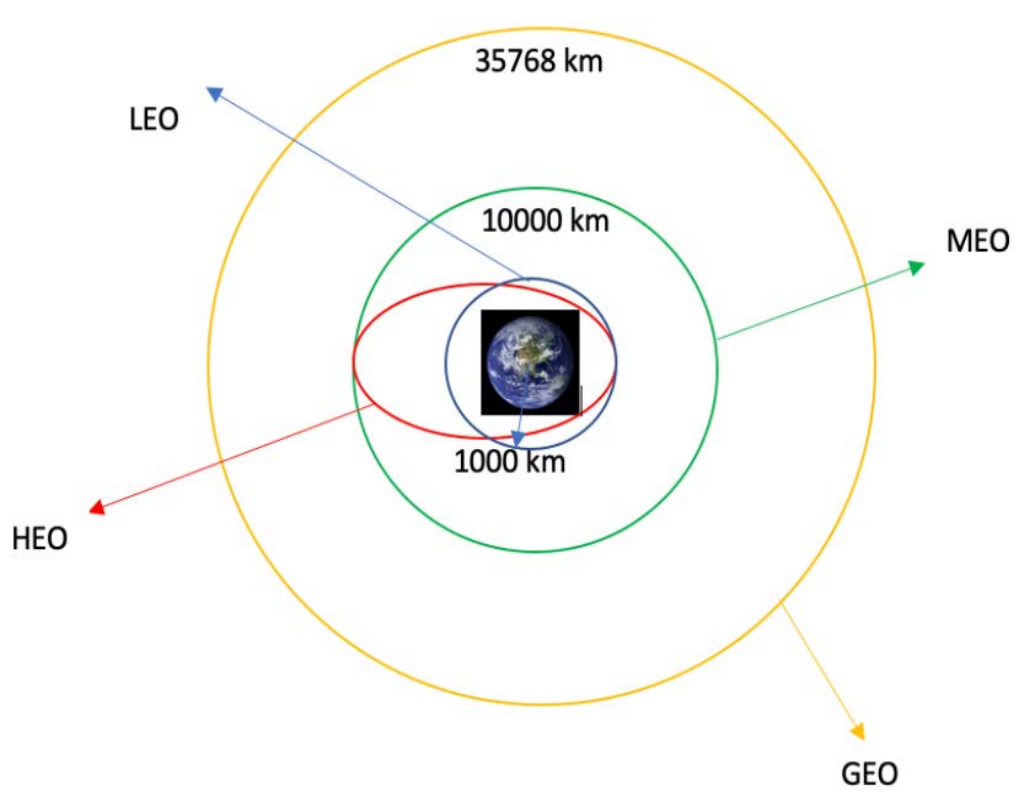
- Data shown that CCAFS SLC 40 is the most preferred site. It was the first launch site and also the first to have a successful landing of the rocket.
- VAFB SLC 4E is the less preferred launch site, despite is god percentage of sucessull recoveries.

Payload vs. Launch Site

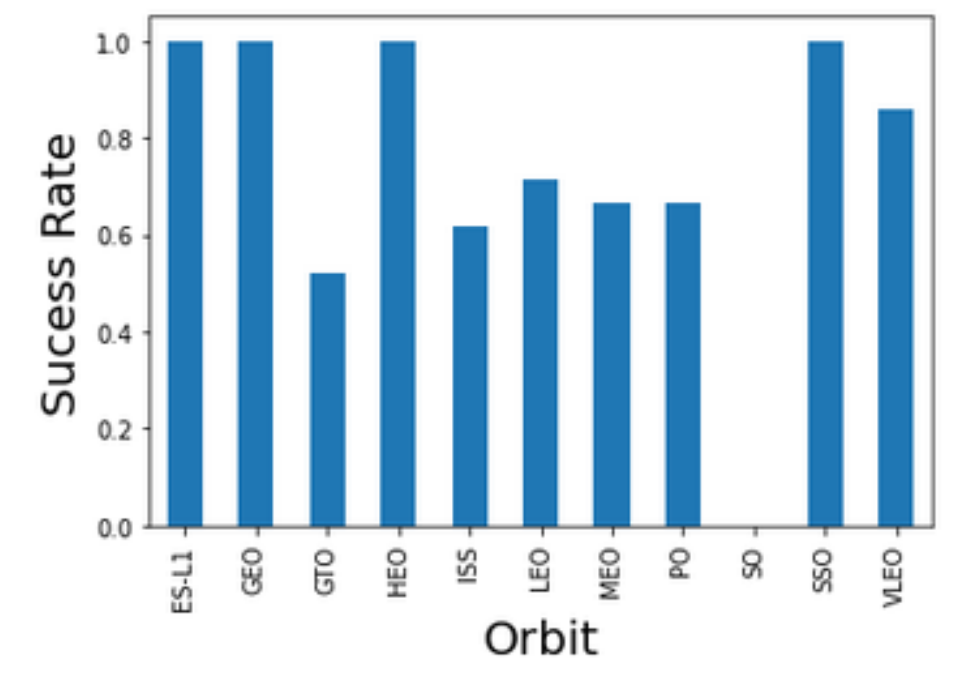


- Data shown that VAFB SLC 4E is only used for payloads up to 10000 kg.
- For heavy loads and for payloads less than 7000 kg, CCAFS SLC 40 is the preferred site.
- KSC LC 39A is used for medium to heavy payloads.

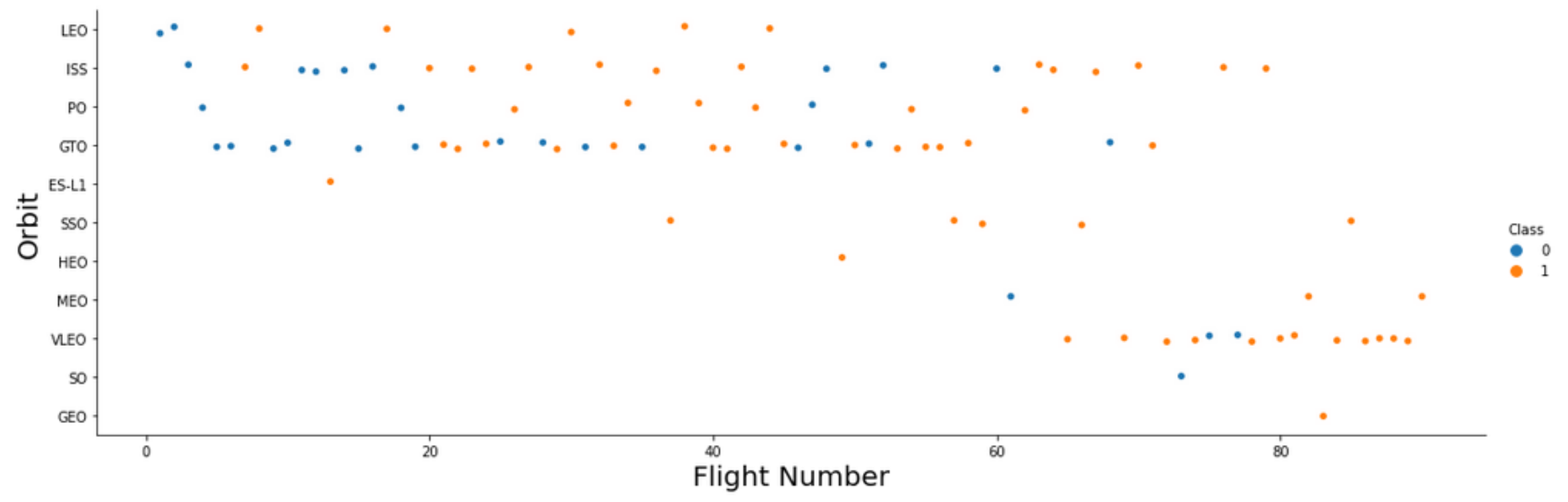
Success Rate vs. Orbit Type



- Data shown that GEO, HEO, SSO and ES-L1 have higher successful missions.
- Low orbits like LEO, ISS, PO and GTO have low rates, since they were the first flights attempted.

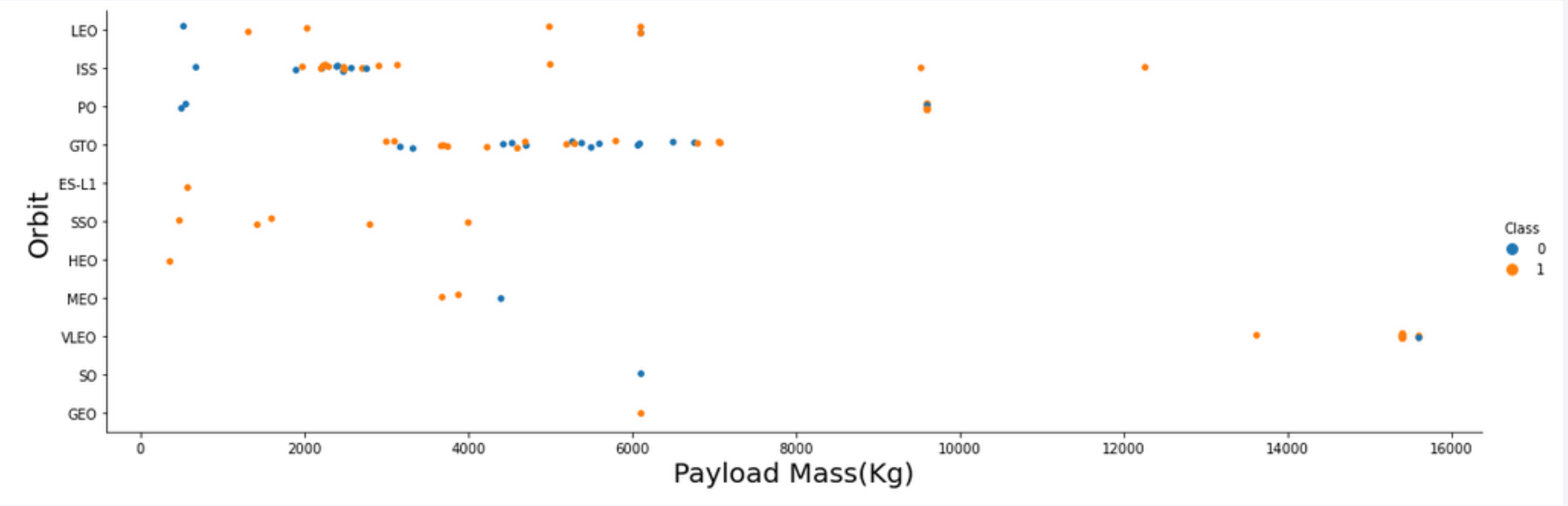


Flight Number vs. Orbit Type



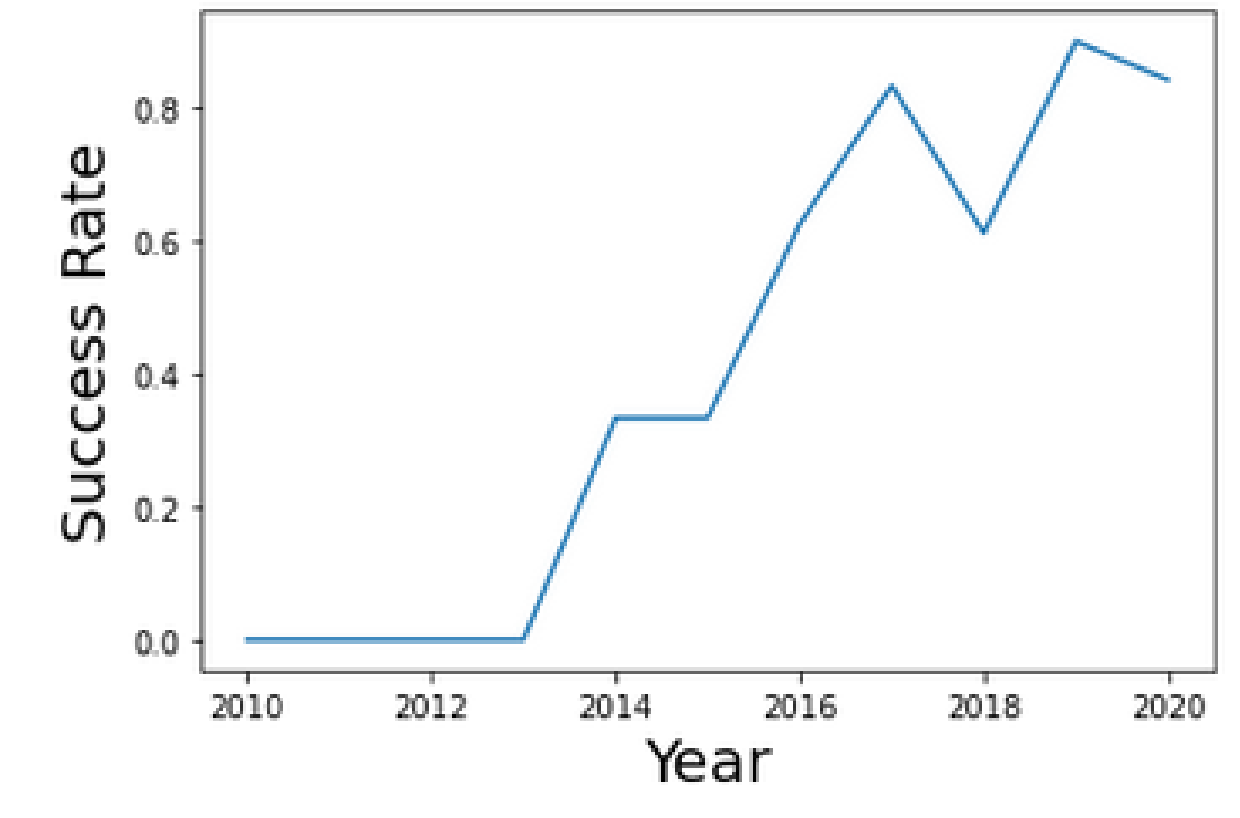
- The first missions attempted lower and transfer orbits (LEO, ISS, PO and GTO). Only after gaining confidence, Space X targeted higher orbits (GEO).
- First successful mission was on a flight to International Space Station.
- For all orbits, as the number of flights increase, success rate also increases.

Payload vs. Orbit Type



- Payloads higher than 13000 kg were deployed only to VLEO (Very low earth orbit), while GEO and SO can only handle payloads lower than 6200 kg.
- Success rate of GTO launches does is not related to payload.

Launch Success Yearly Trend



- All flights after 2014 were unsuccessful. From 2014 to 2017 there was a continuous increase in successful recoveries.
- 2018 was atypical and showed a slight decrease in success rate.

All Launch Site Names

```
In [10]: %sql select distinct(Launch_site) from SPACEXTBL limit 5
```

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

- Query used present the key word “distinct” to present only “unique” values of Launch sites available in the Space X database.
- Limit 5 was used but it has no impact on the final result, since only 4 sites were found in the database.

Launch Site Names Begin with 'CCA'

```
In [25]: %sql SELECT * from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
```

Out[25]:

DATE	time__utc_	booster_version	launch_site
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40

- Query used present the key word “limit” to present the five first records available in the Space X database, with name begins with “CCA”.
- The wildcard was used to allow any combination after the initial letters “CCA”.

Total Payload Mass

```
In [29]: %sql select sum(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL where CUSTOMER LIKE '%NASA (CRS)';
```

payloadmass
45596

- Query used present the key word “sum” to accumulate all payload values that have “Nasa (CRS)”as customer.
- The wildcard was used to allow any initial combination of characters before letters “Nasa”.

Average Payload Mass by F9 v1.1

```
In [30]: %sql select AVG(PAYLOAD_MASS__KG_) as payloadmass from SPACEXTBL where BOOSTER_VERSION LIKE '%F9 v1.1%';
```

payloadmass
2534

- Query used present the key word “average” to calculate the average of all payload values that have a “F9 v1.1” rocket as booster.
- The wildcard was used to allow any initial or final combination of characters before/after letters “F9 v1.1”.

First Successful Ground Landing Date

```
In [18]: %sql select min(DATE) from SPACEXTBL where LANDING__OUTCOME = 'Success (ground pad)';
```

1
2015-12-22

- Query used present the keyword “min” to select the minimum value from all missions that had a successful landing outcome in a ground pad.
- We don´t used wildcards since we want a perfect match of the landing outcome.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
In [31]: %sql select BOOSTER_VERSION from SPACEXTBL where LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

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

- Query used two statements under the where clause.
- The first selected only successful landings in the drone ship and, the second, selected among those results, the booster versions that matched the desired payload.

Total Number of Successful and Failure Mission Outcomes

```
In [36]: %sql select mission_outcome, count(mission_outcome) from SPACEXTBL GROUP BY mission_outcome;
```

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- Query used two keywords in the select statements.
- The first selected the mission outcome status and the second accumulated the values.
- The keyword “Group by” sorted the results by status in a “table” format.

Boosters Carried Maximum Payload

```
In [75]: %sql select booster_version, payload_mass__kg_ from SPACEXTBL where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXTBL)
```

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

- Query used a subquery to filter the boosters that delivered the maximum payload.
- The results were displayed in two columns, the first for the booster number and the second for the payload delivered. All of them were next satellites.

2015 Launch Records

```
%sql select booster_version, launch_site from SPACEXTBL where landing__outcome = 'Failure (drone ship)' and date < '2015-12-01' and Date > '2015-01-01'
```

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- Query used two statements under the where clause.
- The first selected only unsuccessful landings in the drone ship and, the second, selected among those results, those whose happend between 01/January and 31/december/2015.

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

```
In [83]: %sql select count(landing__outcome), landing__outcome from SPACEXTBL \
         where DATE between '2010-06-04' and '2017-03-20' group by landing__outcome\
         order by count(landing__outcome) desc
```

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

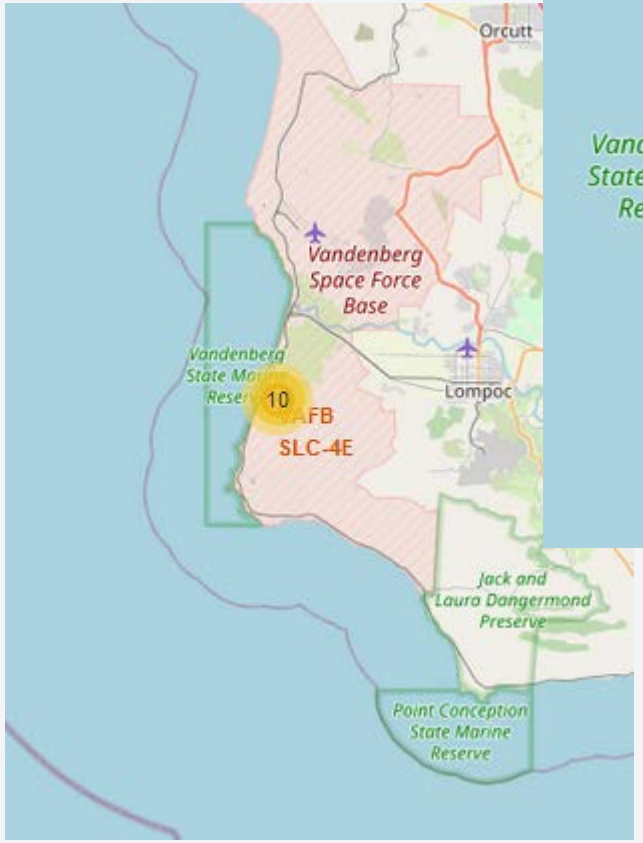
- Query used one statements under the where clause. It selected all landing outcomes between 04/Jun/2010 and 20/Mar/2017.
- Keywords under “select” statement get landing outcomes and accumulate the values.
- Keywords Group by and order are used to show the total per outcome type and to sort then in a descending order.

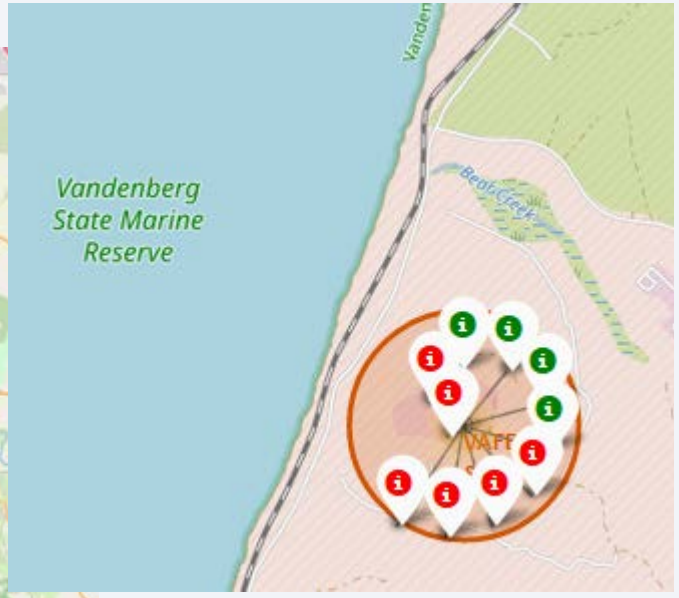

Section 3
Launch Sites
Proximities Analysis

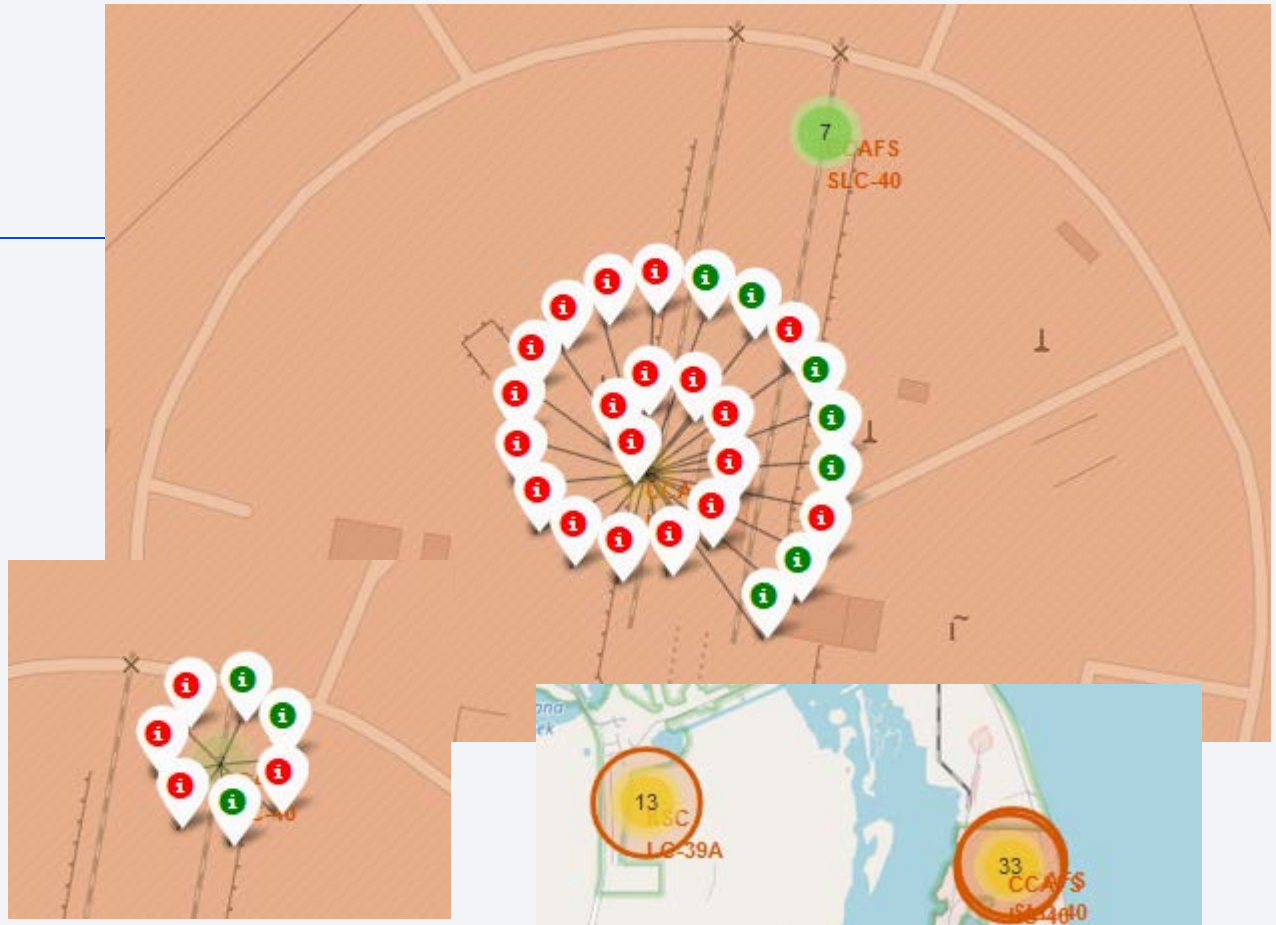
Space X Launch Sites

- All Space X launch sites are located near the coastline and Equator line.

Success rate

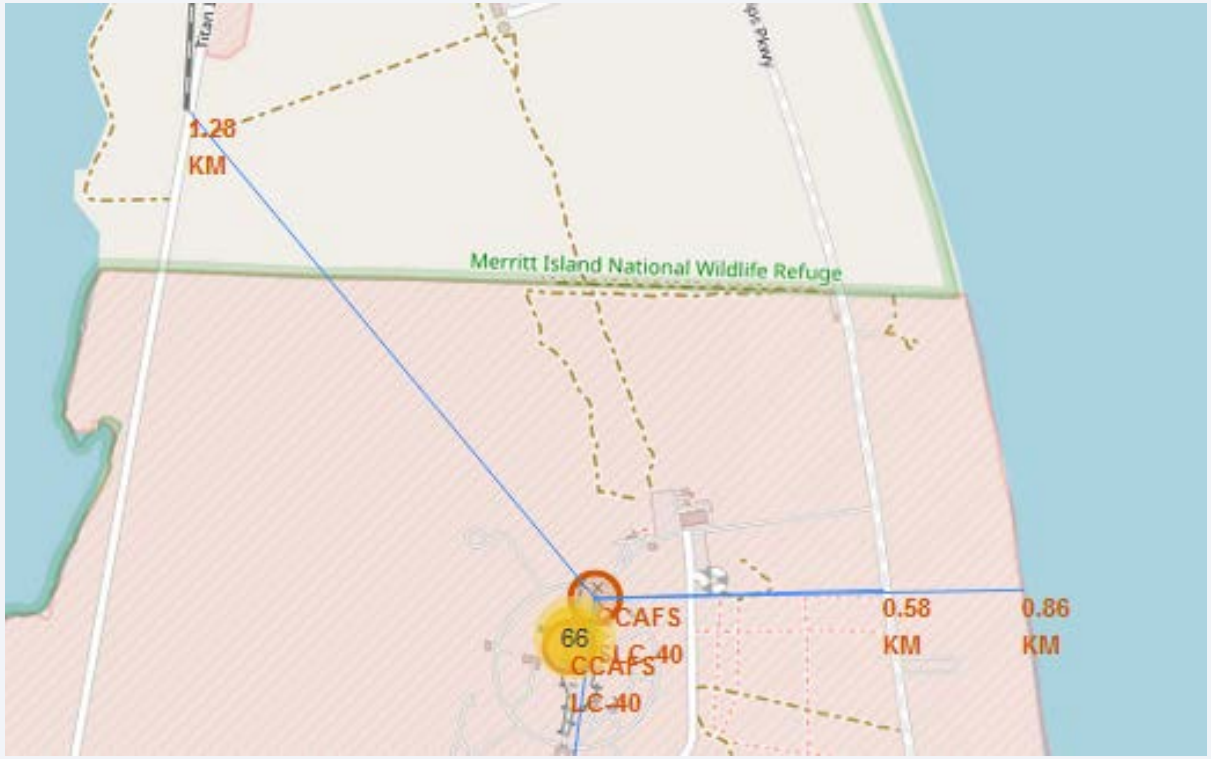


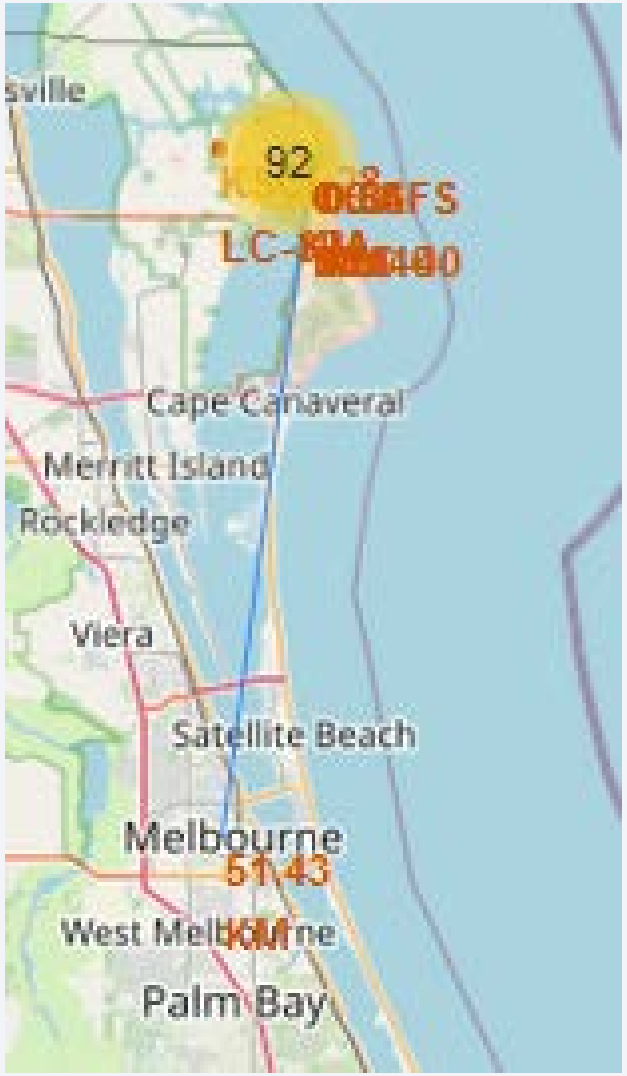




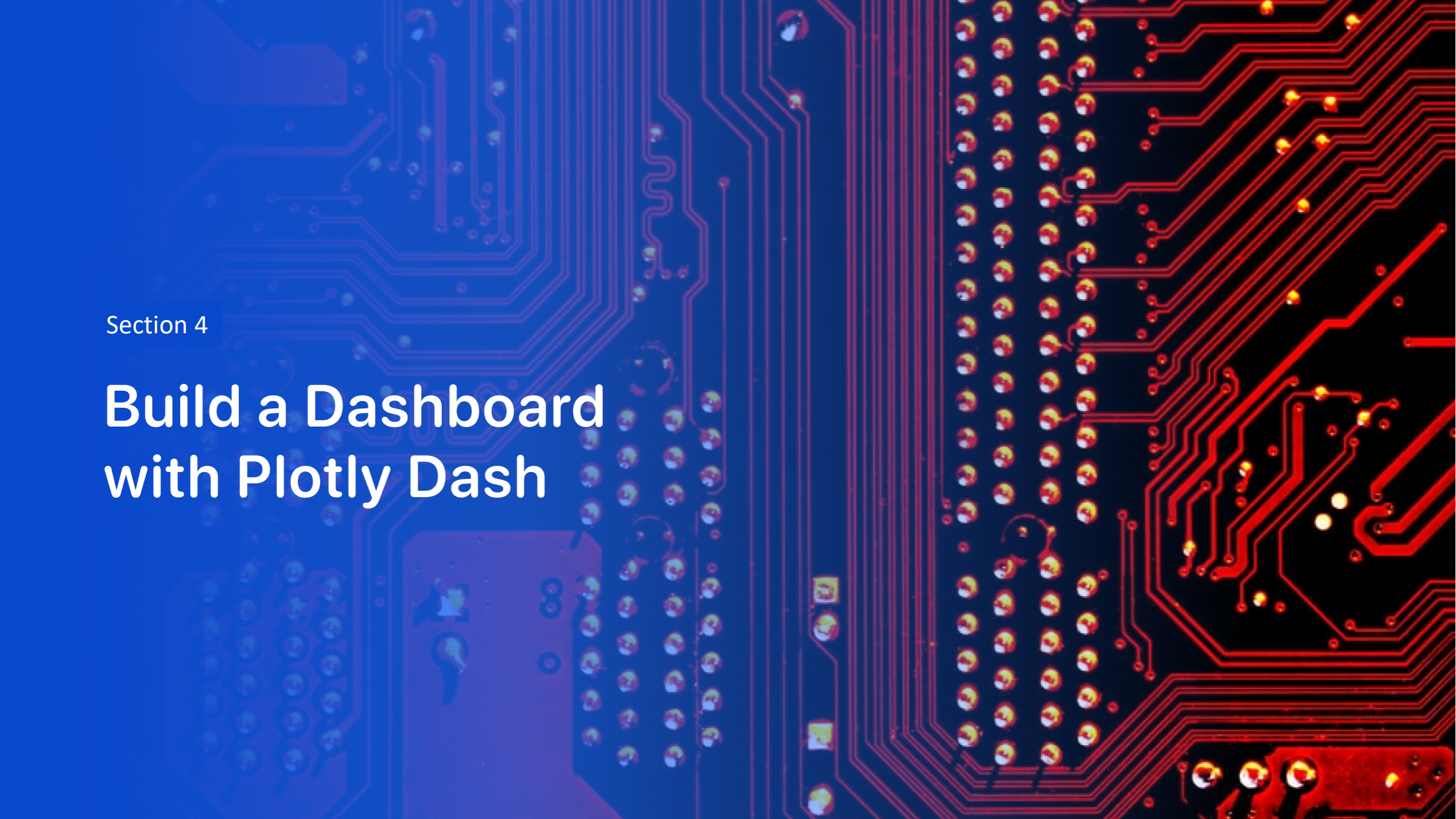
- Most flights departed from Florida and CCAFS SLC 40 launch site.

Neighborhood

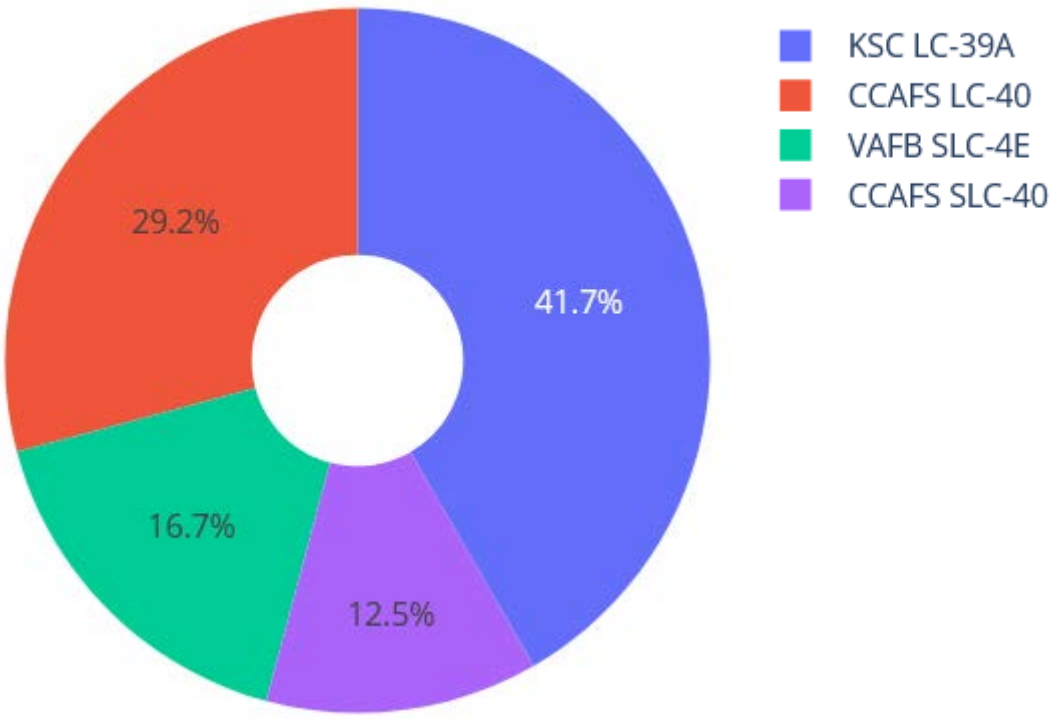




- Launch sites are closer to coastline and railway, but far from major cities.
- Railway is important to bring supplies and the coastline allows a safe exit to rockets in case of a malfunction during launch.


Section 4
Build a Dashboard with Plotly Dash

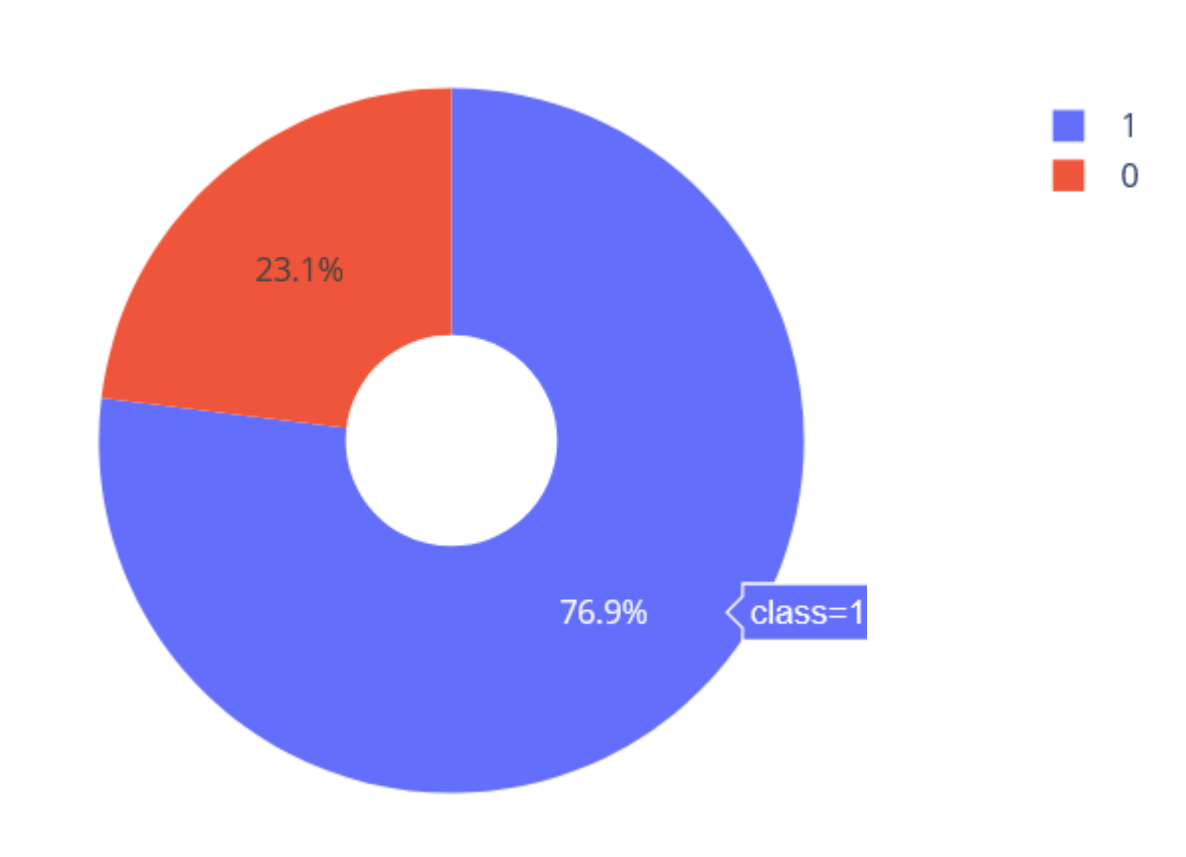
Space X launch sites



- Data shown that CCAFS SLC 40 has the lower success rate among all launch sites. It has only 12,5 % of success rate.

- On the other side, KSC LC 39A has 41,7 % of success rate, followed by CCAFS LC-40 with 29,2 %.

Site matters



- Drilling in data for KSC LC 39A shows that 76,9 % of mission from this pad had a positive outcome.

Payload matters

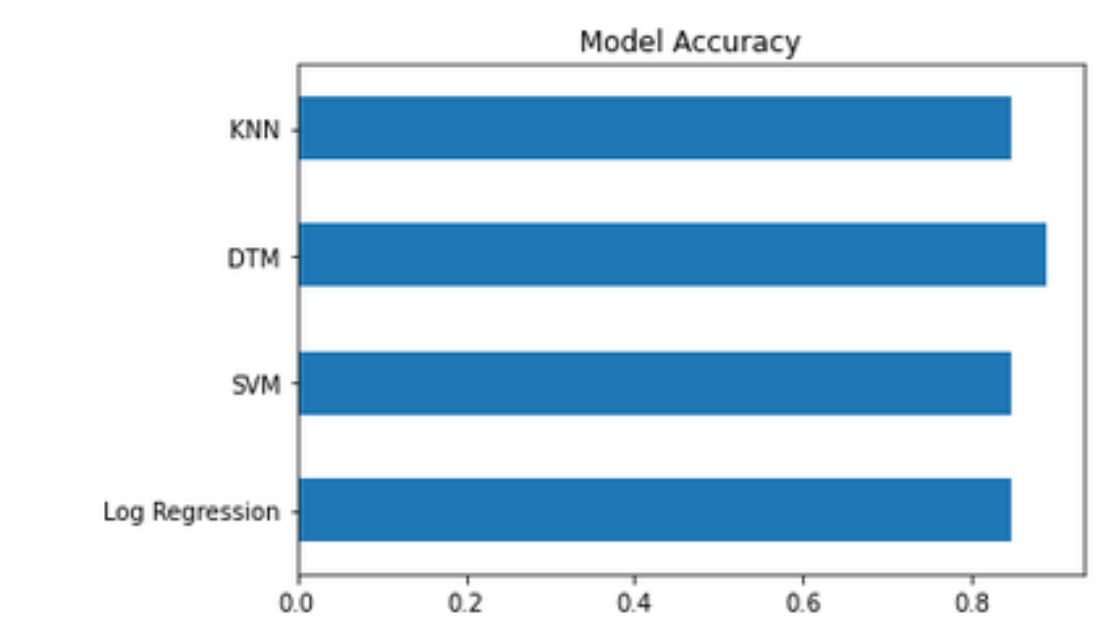
- Data has shown that payloads are a key factor for successful recovery of boosters.

- Payloads higher than 6000 kg present a very low successful rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



- The bar chart on the left shows the accuracy for all models. Accuracy for all models was quite similar:
  - Logistics Regression method: 0.8464
  - Support Vector Machine method: 0.8482
  - Decision tree method: 0.8875
  - K nearest neighbors method: 0.8482
- Decision tree model present a slightly better accuracy and was appointed as the chosen one.

Confusion Matrix

- All confusion matrix were equal.
- All models predicted three landings for mission thar ended with an unsuccesfull outcome.

Conclusions

- Space X only got success after many losses and unsuccessful booster recoveries.
- After 2013 successful rate increased steadily until 2018.
- Higher orbit were only attempted after some confidence was gained with lower orbit launches.
- ES-L1, GEO, HEO, SSO and VLEO orbits had higher successful rate.
- KSC LC 39A has 41,7 % of success rate. Launch sites in Florida are preferred.
- The Decision tree classifier is the best machine learning algorithm, with accuracy slightly better than Logistic Regression, SVM or KNN.

Thank you!