Winning Space Race with Data Science

Name – Debabrata Garai
Date – 27-12-2021

Table of Contents

Executive Summary

Project Details :-

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards of 165 million dollars each, much of savings is due to the reusability of first stage. SpaceX's Falcon 9 can recover the first stage. Sometimes it does not land or sometimes it crashes or in other times, SpaceX sacrifices the first stage due to the mission parameters like payload, orbit and customer. SpaceY is a new rocket company which would like to compete with SpaceX founded by Billionaire industrialist Allon Musk.
- Our job is to find the correlation between the first stage landing success rate and launch parameters like payload mass, orbit type, booster types, where it is launched from and more. Therefore if we can determine if the first stage will land, we can determine the cost of launch for SpaceY.
- We used the SpaceX Rest API and web scraping to gather information available publicly and machine learning to predict if SpaceX will reuse the first stage. We also performed Exploratory Data Analysis and Data Visualization to gather insights from our data.

Introduction

Problem Statement :-

We want to determine if the first stage will land so that we can estimate the cost of each launch for SpaceY

Nature of Analysis :-

The nature of our analysis was Exploratory as well as Predictive. We performed EDA on the data we gathered and use data visualization to find correlations and prove our hypotheses. Then we used predictive modelling to predict our outcome.

Methodology :-

We gathered information about SpaceX using APIs and web scraping. We created dashboards with Tableau. We also created an interactive map with Folium to locate launch sites and have an idea about the geographical data. Finally we trained a machine learning model and use public information to predict if the first stage will land.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX open-source Rest API for launch, rocket, core, capsule, starlink, launchpad and landing pad data
  - Web Scraping related Wiki pages with some HTML tables that contain valuable Falcon 9 launch records with python BeautifulSoup package
- Perform data wrangling
  - Wrangling Data using an API
  - Sampling Data
  - Dealing with Nulls

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

  We will test:-

  - Logistic Regression
  - Support Vector machines
  - Decision Tree Classifiers
  - K-nearest neighbors

  We will find the hyperparameters that allow the given algorithm to perform best.

Data Collection

Data was collected in two steps:-

- Using SpaceX open-source Rest API for launch, rocket, core, capsule, starlink, launchpad and landing pad data

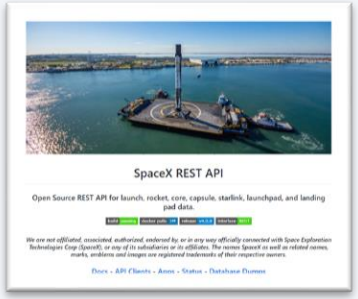
URL – 'https://api.spacexdata.com/v4/launches/past'

We performed a get request using the requests library to obtain the launch data, which will use to get the data from the API.

```
response = requests.get(url)

response.json()

dataframe = pd.json_normalize(response.json())
```



- Web Scraping related Wiki pages with some HTML tables that contain valuable Falcon 9 launch records with python BeautifulSoup package

We parsed the data from those tables and converted them into a Pandas data frame for further visualization and analysis. For some specific columns like Booster version, launchpad, payload and core we had an identification number and no actual data. So, we used the API again targeting another endpoint to gather specific data for each ID number.

Data Collection – SpaceX API

- We used different API calls targeting endpoints corresponding to rocket, cores, launchpad and payloads columns to extract information using each ID number in the launch data
  - api.spacexdata.com/v4/rockets
  - api.spacexdata.com/v4/cores
  - api.spacexdata.com/v4/launchpads
  - api.spacexdata.com/v4/payloads
- GitHub URL of the SpaceX API calls notebook :-

https://github.com/devoeop/IBM_capstone/blob/main/Data%20Collection%20API.ipynb

We defined a series of helper functions that uses the API to extract information using ID numbers

After collecting the raw data which was stored in lists, we created our dataset.

After that we sampled the data to remove Falcon 1 launches.

We also removed the null values in order to make the dataset viable for analysis.

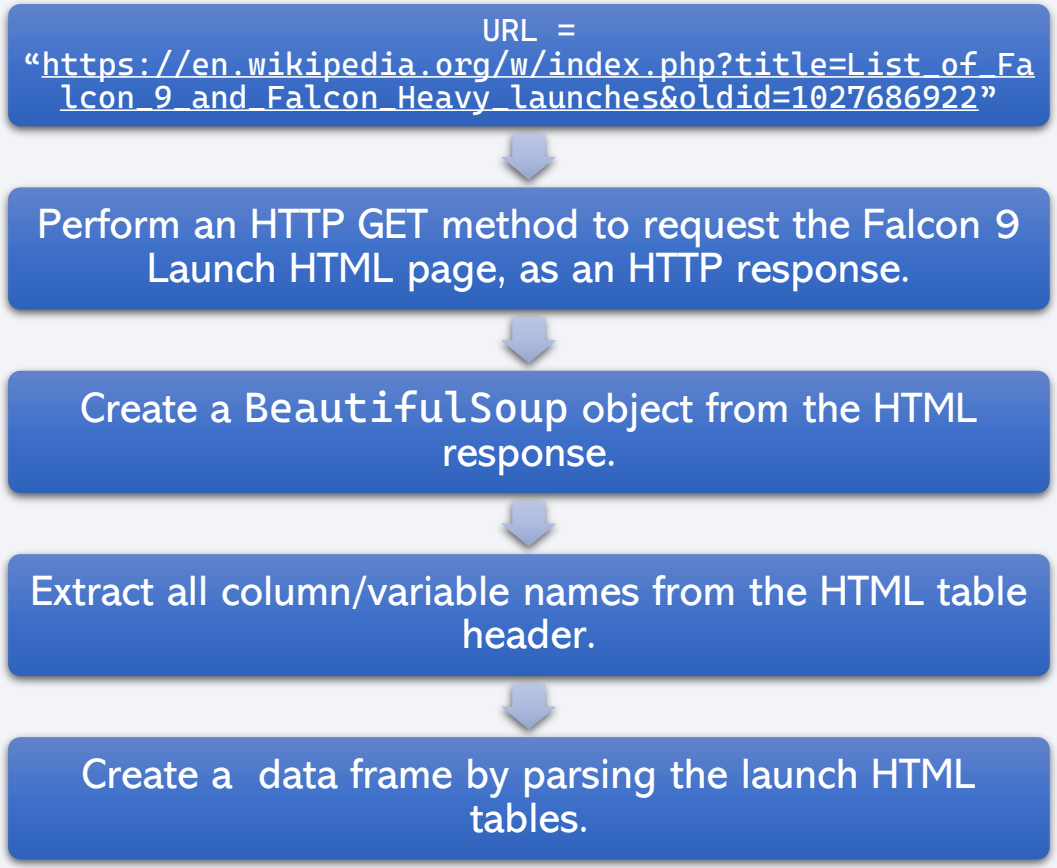
We replaced the null values in PayloadMass with the mean of the PayloadMass data.

We left the column LandingPad with NULL values as it represented when a landing pad is not used.

Data Collection - Scraping

- Web scraped Falcon 9 launch records with `BeautifulSoup`:
  - Extracted a Falcon 9 launch records HTML table from Wikipedia
  - Parsed the table and convert it into a data frame
- GitHub URL of the SpaceX API calls notebook :-

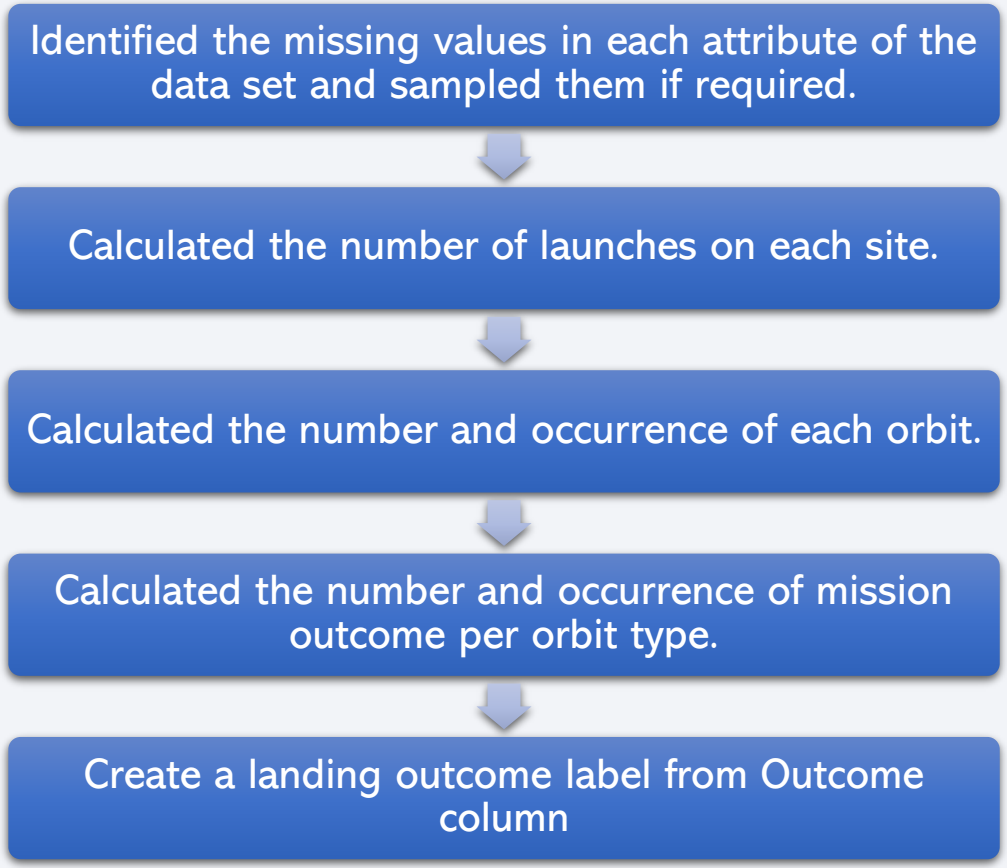
https://github.com/devoeop/IBM_capstone/blob/main/Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

- We performed some EDA to find some patterns in the data and determine what would be the label for training supervised models.
  - There were 8 types of outcomes. [True ASDS, None None, True RTLS, False ASDS, True Ocean, None ASDS, False Ocean, False RTLS]
  - 3 were labelled successful outcomes(landing_class = 1).
  - 5 were labelled bad outcomes(landing_class = 0)
- GitHub URL of the data wrangling notebook :-

https://github.com/devoeop/IBM_capstone/blob/main/Data%20Wrangling%20-%20EDA.ipynb



EDA with SQL

- We loaded the data set into the corresponding table in a Db2 database
- We then executed SQL queries to gather insights about the data set
  - Querying the names of the unique launch sites in the space mission and their records.
  - Querying about the payload mass metrics carried by specific boosters.
  - Querying dates when successful landing outcomes were achieved.
  - Querying names of the boosters which have success in drone ship with specific payload mass.
  - Querying total number of successful and failure mission outcomes.
  - Querying failed landing outcomes.
- GitHub URL of the EDA with SQL notebook :-

https://github.com/devoeop/IBM_capstone/blob/main/EDA%20with%20SQL.ipynb

EDA with Data Visualization

- Performed EDA and Feature Engineering using `Pandas` and `Matplotlib` to determine if the first stage can be reused.
  - Using a line chart, we plotted the launch success rate.
  - Using a bar plot, we saw that different launch sites have different success rates. As a result, they can be used to help determine if the first stage will land successfully.
  - We plotted scatterplots of different metrics to find trends.
  - Using a bar plot, we found the success related to each orbit type
  - Created dummy variables for categorical columns and applied OneHotEncoding
- GitHub URL of the EDA with data visualization notebook :-

https://github.com/devoeop/IBM_capstone/blob/main/EDA%20with%20Data%20Visualization.ipynb

Build an Interactive Map with Folium

- Created an interactive map with the `Folium` library.
  - Marked all launch sites on the map with `folium.Circle` and `folium.Marker`. We found that most of the launch sites were within 1.5km distance from the coastline.
  - Marked the successful and failed launches for each site on the map. We used Marker clusters to show launch records having the exact same coordinate.
  - Calculated and marked the distances between launch site and its proximities. Added a `MousePosition` to get coordinates of any point of interests using a mouse over a point on the map.
  - Draw a `Polyline` between launch sites and its proximities.
- GitHub URL of the Interactive Visual Analytics with Folium notebook :-

https://github.com/devoeop/IBM_capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Tableau

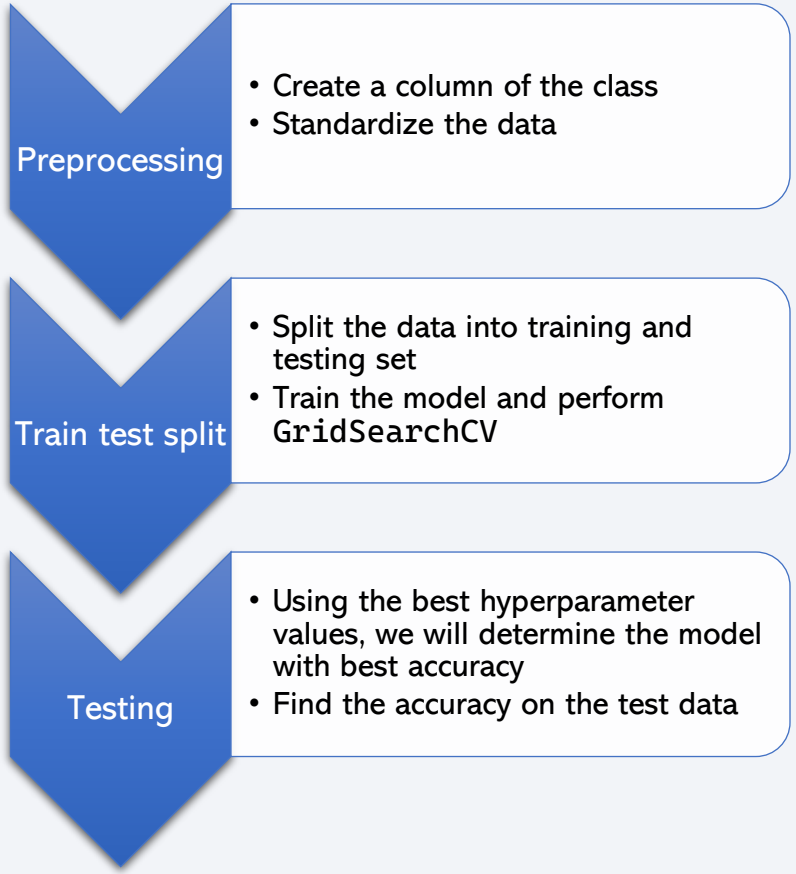
- Visualizations :-
  - Added a pie chart showing Total success rate by Site
  - Added a pie chart with dropdown menu of launch sites showing Launch success rate
  - Added a scatterplot with range slider of payload mass showing launch outcome vs payload mass
- URL of the Dashboard :-

https://public.tableau.com/views/SpaceXLaunchRecordsDashboard/SpaceXLaunchRecordsDashboard?:language=en-US&:display_count=n&:origin=viz_share_link

Predictive Analysis (Classification)

- We built a Machine Learning pipeline to predict if the first stage of the Falcon 9 land successfully.
  - We preprocessed the data and determined training labels
  - Using the function `train_test_split` to split the data X and Y into training and testing data in the ratio 8:2
  - Create `Logistic Regression, SVM, Decision Tree Classifier` and `K nearest neighbors` object
  - `GridSearchCV` and fit the object to find the best hyperparameters for each model
  - Find the accuracy of each model on the test data
  - Find the method performs best using test data
- GitHub URL of the predictive analysis lab notebook :-

https://github.com/devoeop/IBM_capstone/blob/main/Machine%20Learning%20Prediction.ipynb

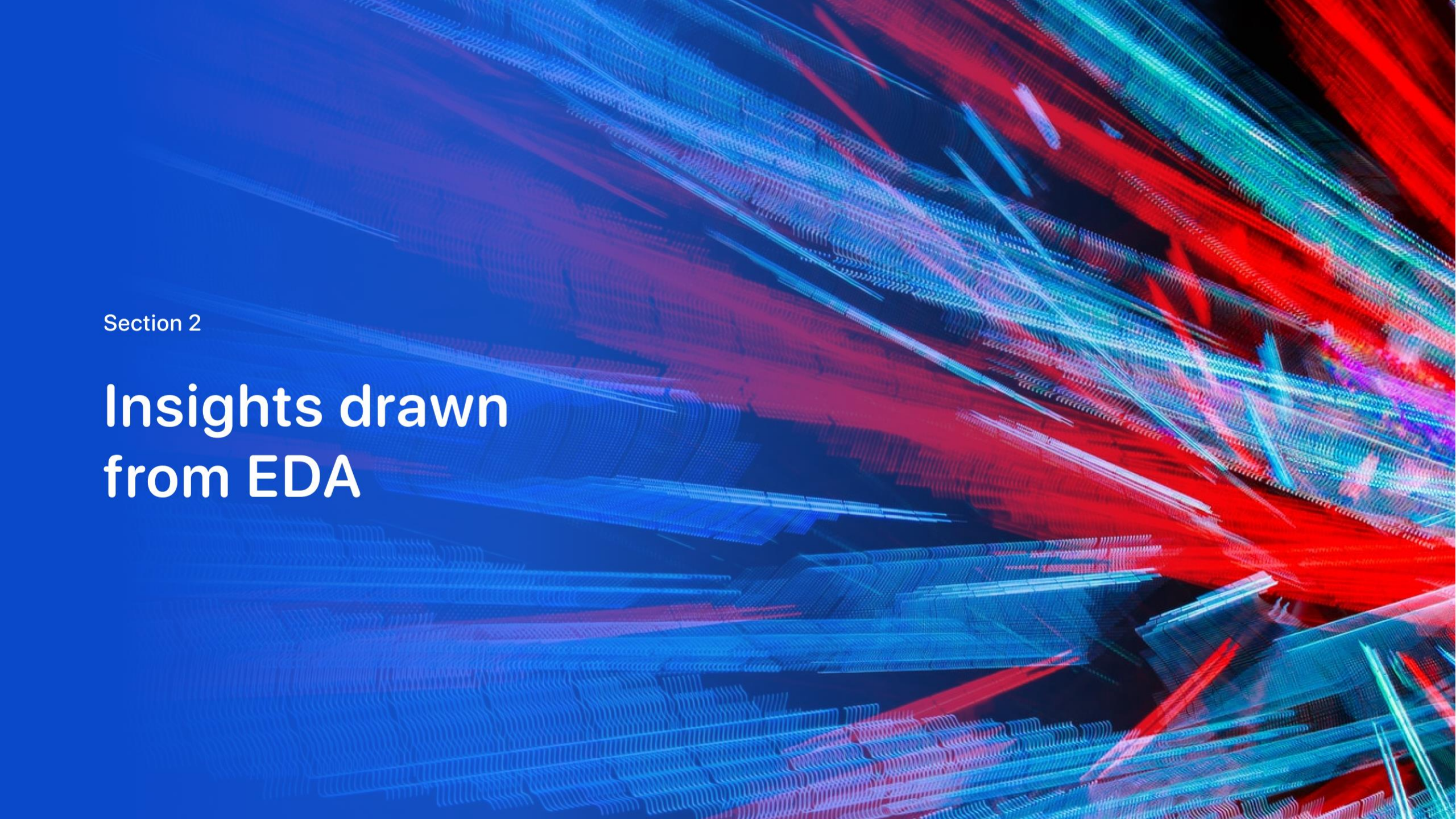


Results

- The Data we collected was relatively small for our analysis but since there were no other data available, we used these datasets only. Our dataset consisted of one 90 rows x 17 columns collected from rest API and another one of 121 rows and 11 columns scraped from web
- We created a Class label based on Landing Outcome where True ASDS, True RTLS, True Ocean were considered successful outcomes and None None, False ASDS, None ASDS, False Ocean, False RTLS were considered bad outcomes. Successful outcomes were labelled 1 and bad outcomes were labelled 0
- We created a dashboard using Tableau and found KSC LC-39A had the best launch success rate
- Our best Machine Learning model had an accuracy of about 89%

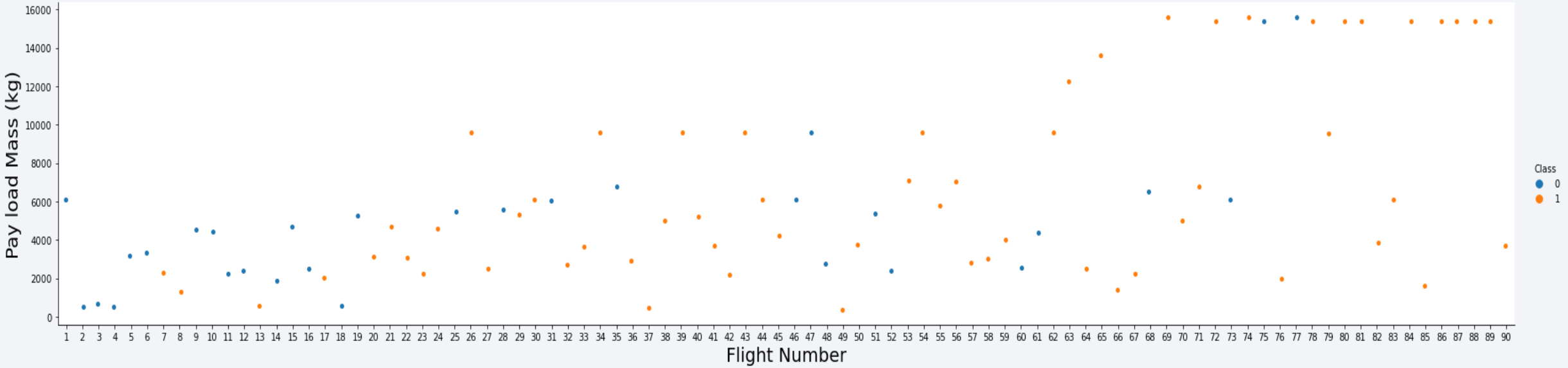
Section 2

Insights drawn from EDA



Flight Number vs. Payload

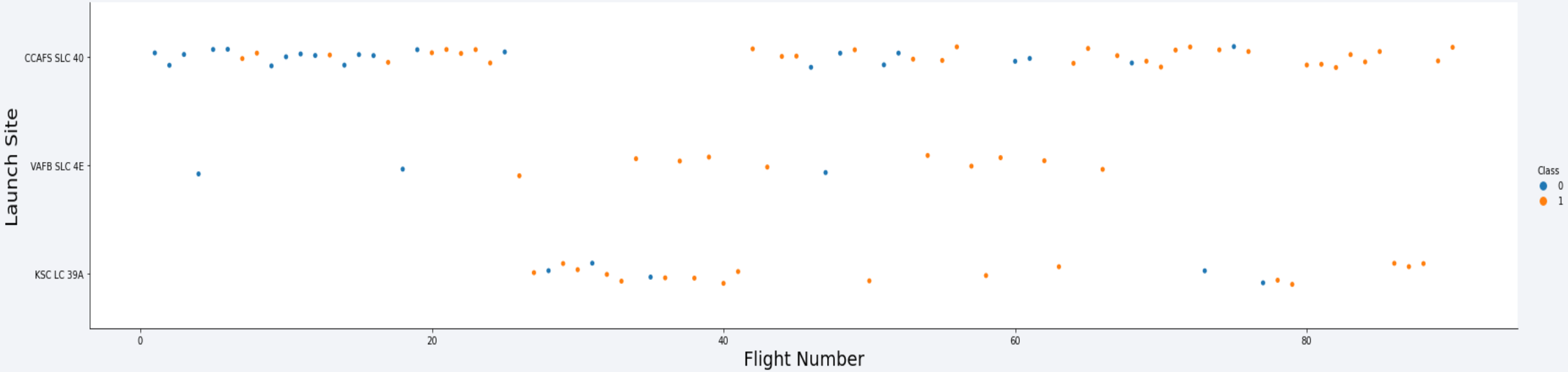
Scatter plot of Flight Number vs. Payload



We can see that as the Flight Number(indicating continuous launch attempts) increases, the first stage is more likely to land successfully. Also the more massive the payload, the less likely the first stage will return.

Flight Number vs. Launch Site

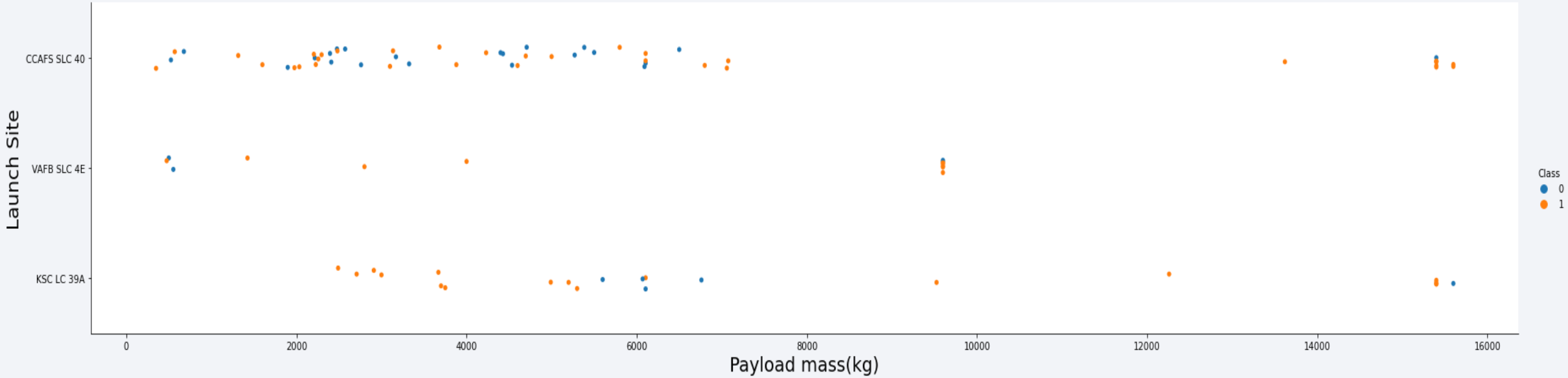
Scatter plot of Flight Number vs. Launch Site



We can see from the graph that there is no definitive relation between Flight Number and launch Site although we can say that launch site Vandenberg Air Force Base Space Launch Complex (VAFB SLC 4E) has more successful cases when the flight number increases

Payload vs. Launch Site

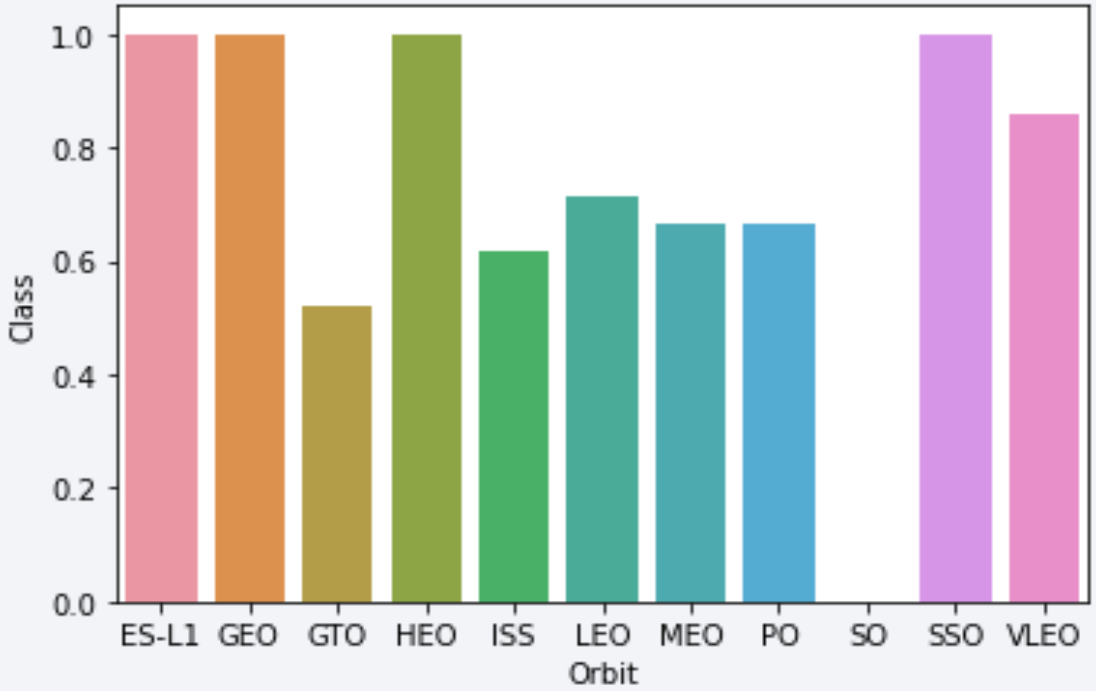
Scatter plot of Payload vs. Launch Site



If we observe Payload Vs. Launch Site scatter point chart, we will find for the Vandenberg Air Force Base Space Launch Complex (VAFB SLC 4E) launchsite there are no rockets launched for heavy payload mass(greater than 10000).

Success Rate vs. Orbit Type

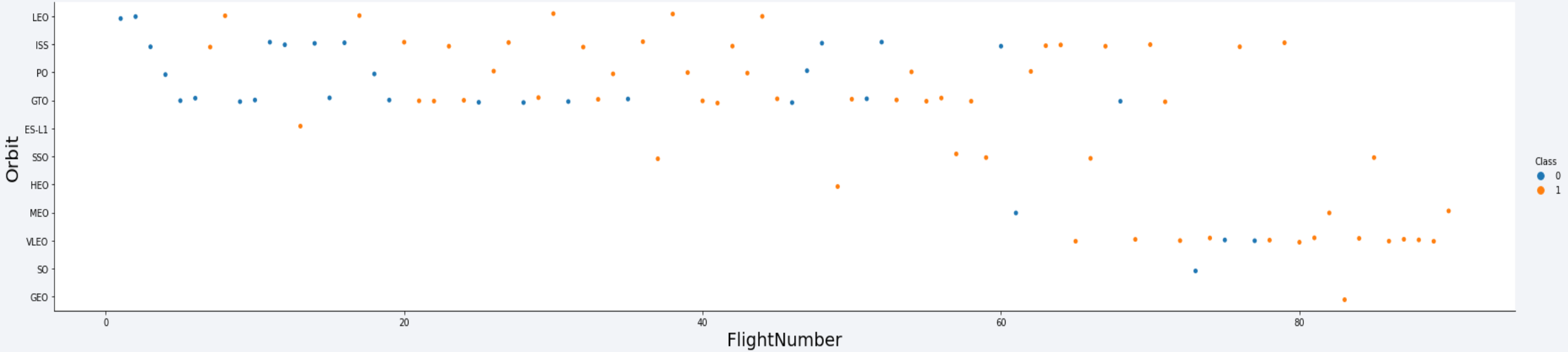
Bar chart for the success rate of each orbit type



We can see that ES-L1, GEO, HEO and SSO orbit types have the highest success rates with mean Class value of 1

Flight Number vs. Orbit Type

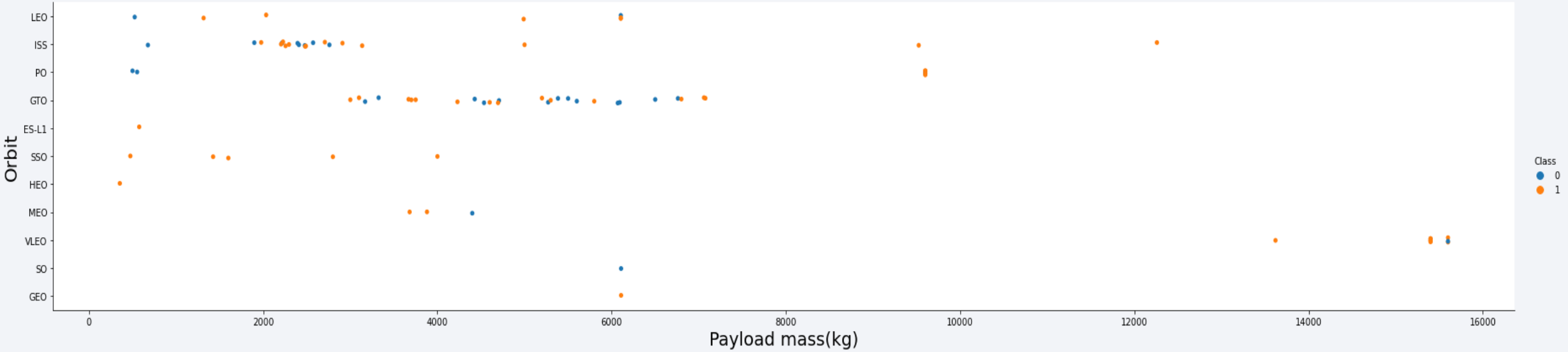
Scatter point of Flight number vs. Orbit type



We can 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.

Payload vs. Orbit Type

Scatter point of payload vs. orbit type

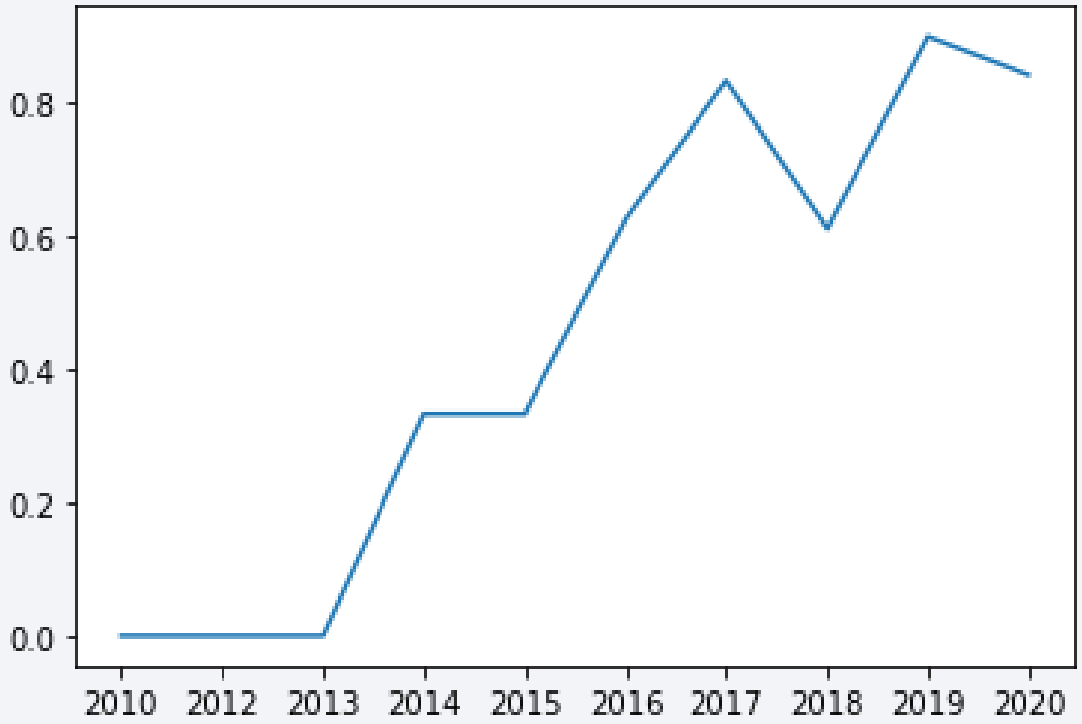


With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.

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

Launch Success Yearly Trend

Line chart of yearly average success rate



We can observe that the launch success rate i.e., the landing of the first stage successfully since 2013 kept increasing till 2020

All Launch Site Names

Names of the unique launch sites

```
%%sql
select distinct launch_site
from VSC30004.SPACEXTBL
```

```
 * ibm_db_sa://vsc30004:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

```
%%sql
select *
from VSC30004.SPACEXTBL
where SUBSTR(LAUNCH_SITE,1,3) ='CCA'
limit 5
```

```
 * ibm_db_sa://vsc30004:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31498/bludb
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	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

The total payload carried by boosters from NASA

```
%%sql
select sum(payload_mass__kg_)
from VSC30004.SPACEXTBL
where substr(customer,1,4) = 'NASA'
```

```
 * ibm_db_sa://vsc30004:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

1
99980

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1

```
%%sql
select avg(payload_mass__kg_)
from VSC30004.SPACEXTBL
where substr(booster_version,1,7) = 'F9 v1.1'
```

```
 * ibm_db_sa://vsc30004:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

1
2534

First Successful Ground Landing Date

Dates of the first successful landing outcome on ground pad

```
%%sql
select min(date)
from VSC30004.SPACEXTBL
where landing__outcome = 'Success (ground pad)'
```

```
 * ibm_db_sa://vsc30004:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%%sql
select booster_version
from VSC30004.SPACEXTBL
where landing__outcome = 'Success (drone ship)' and payload_mass__kg_>4000 and payload_mass__kg_<6000
```

```
 * ibm_db_sa://vsc30004:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
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

Total number of successful and failure mission outcomes

```
%%sql
select count(mission_outcome), mission_outcome
from VSC30004.SPACEXTBL
group by mission_outcome
```

```
 * ibm_db_sa://vsc30004:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

Boosters Carried Maximum Payload

Names of the booster which have carried the maximum payload mass

```
%%sql
select booster_version, payload_mass__kg_
from VSC30004.SPACEXTBL
where payload_mass__kg_ = (select max(payload_mass__kg_) from VSC30004.SPACEXTBL)
```

```
 * ibm_db_sa://vsc30004:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31498/bludb
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

2015 Launch Records

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

```
%%sql
select date, booster_version, launch_site, landing__outcome
from VSC30004.SPACEXTBL
where substr(landing__outcome,1,7) = 'Failure' and year(date) = 2015
```

```
 * ibm_db_sa://vsc30004:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

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

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

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

```
%%sql
select count(landing__outcome) as count, landing__outcome
from VSC30004.SPACEXTBL
where date between '2010-06-04' and '2017-03-20'
group by landing__outcome
order by count(landing__outcome) desc
```

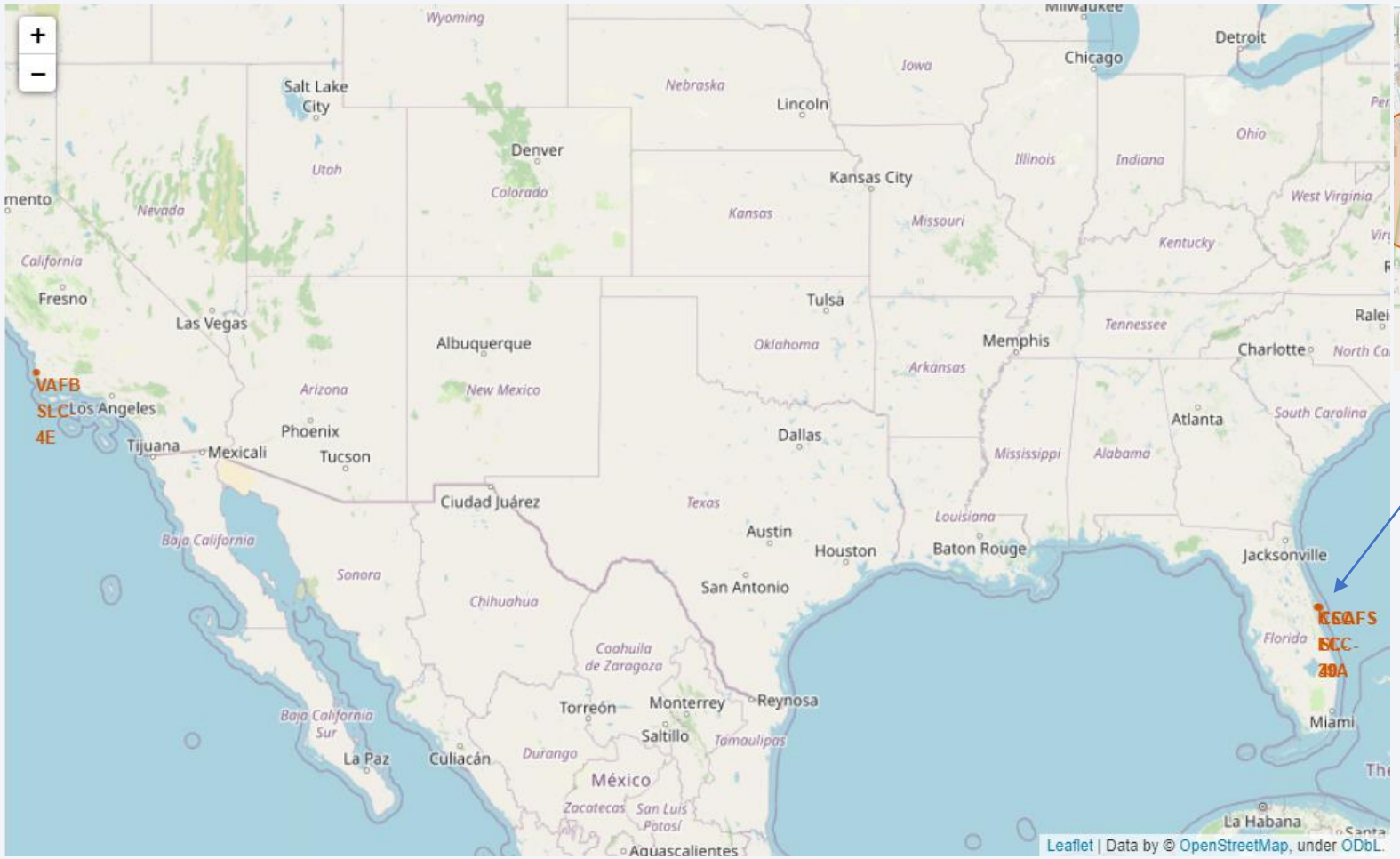
```
 * ibm_db_sa://vsc30004:***@3883e7e4-18f5-4afe-be8c-fa31c41761d2.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31498/bludb
Done.
```

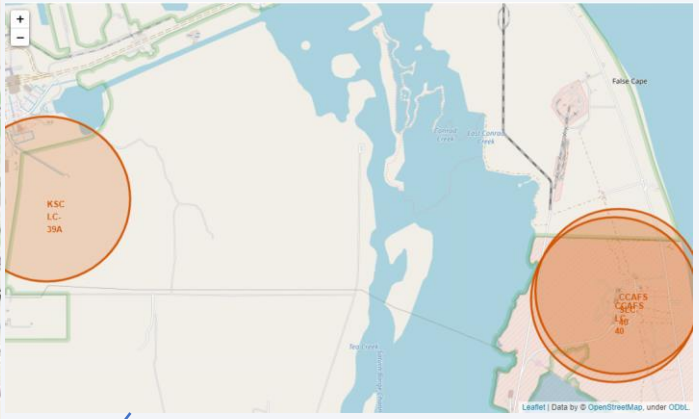
COUNT	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

Section 4

Launch Sites Proximities Analysis

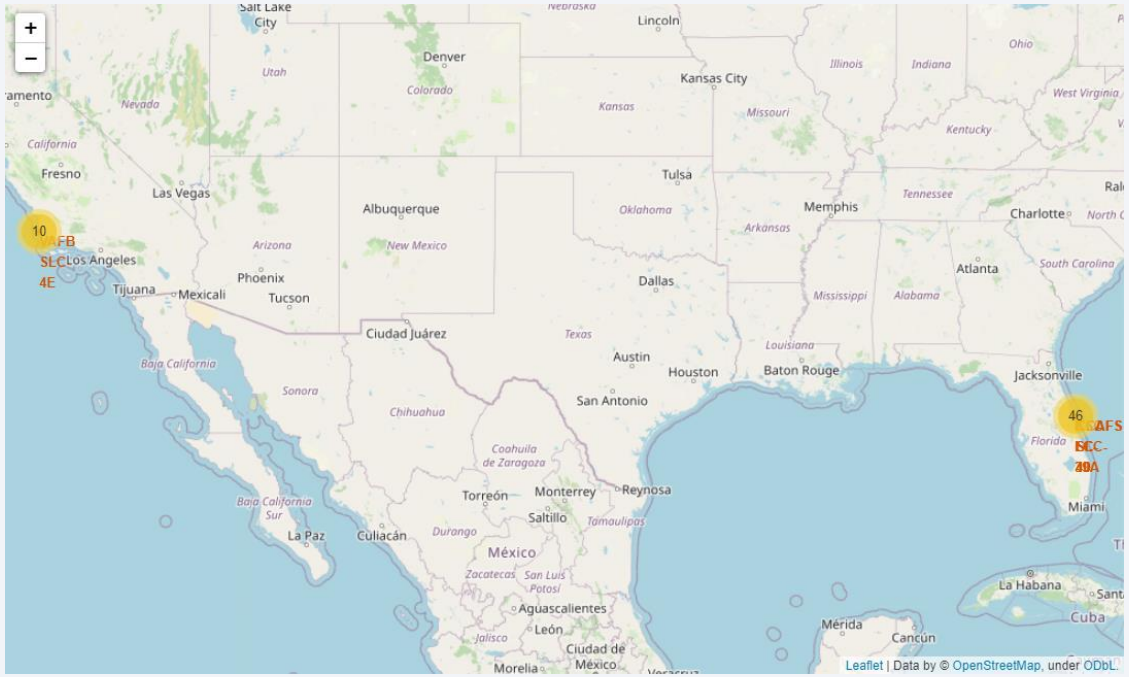
Launch sites Locations





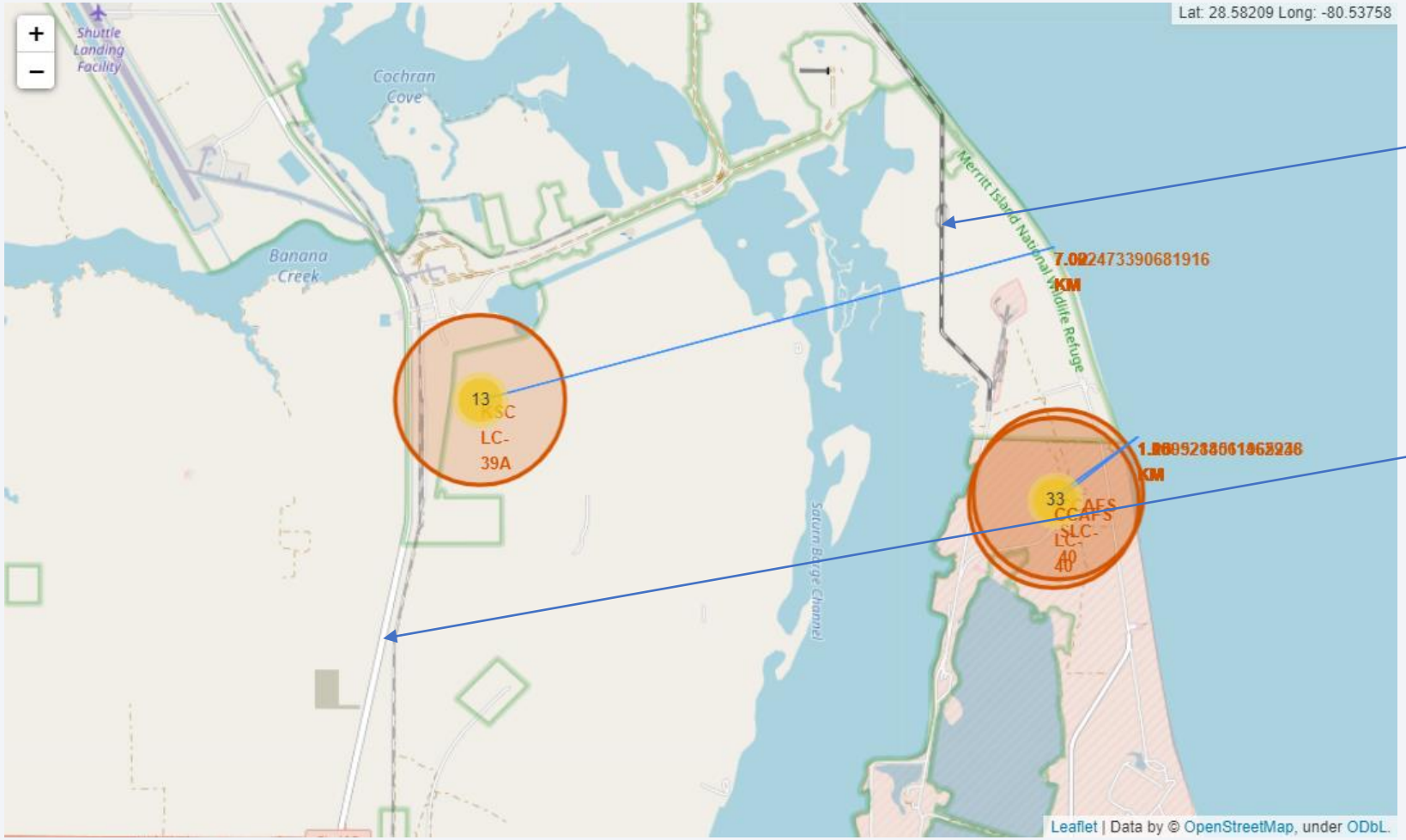
From the picture we can see that there are 4 launch site locations in USA

Launch Outcomes



- The first picture shows marker clusters for the launch records on the launch sites
- The second picture shows the markers for the launch outcomes
- Green marker indicate successful outcome whereas red marker indicate failed outcome

Launch site proximities

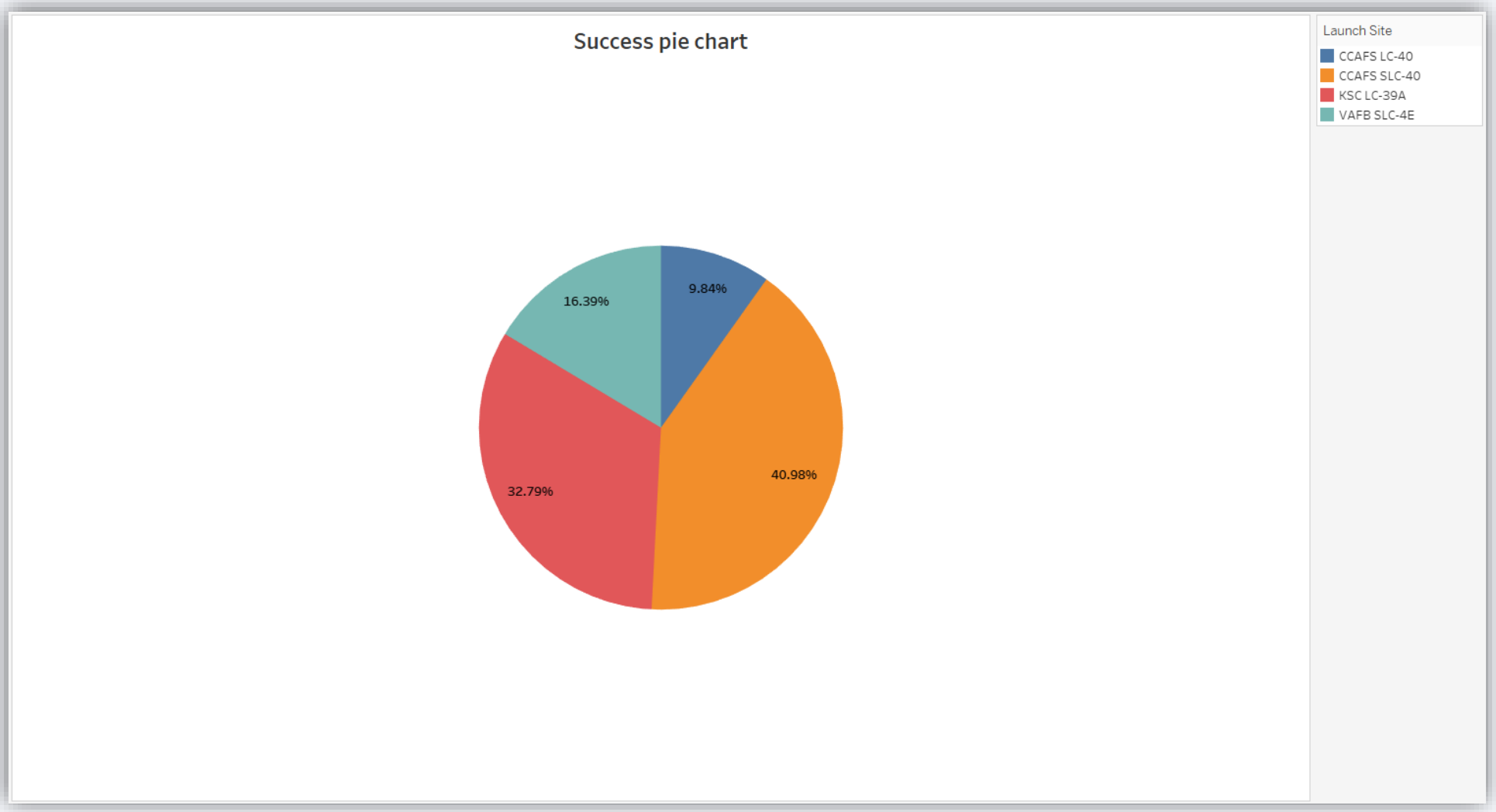


This picture shows the distance between the launch sites and their nearest coastlines.

Section 5

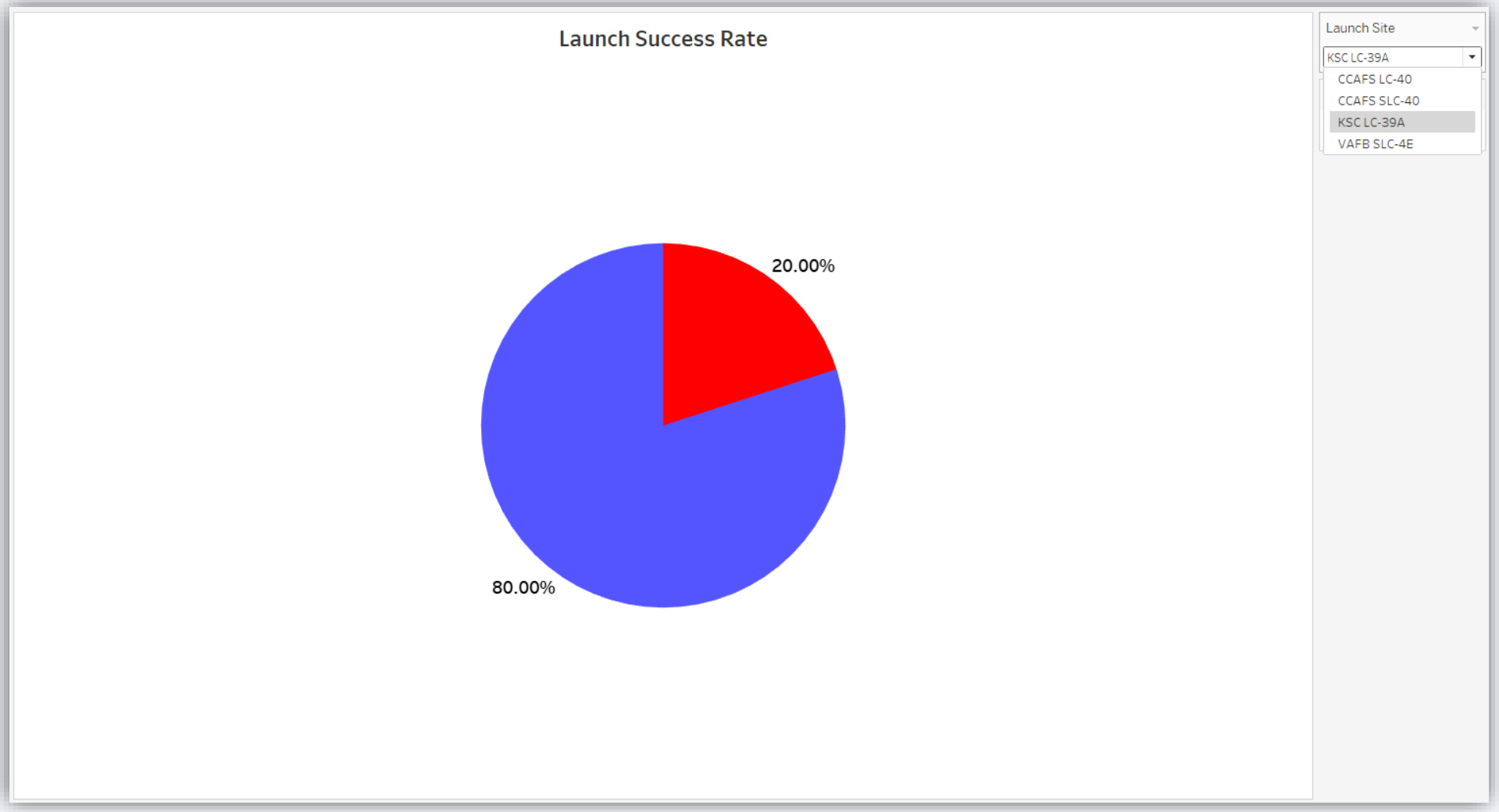
Build a Dashboard with Tableau

Total Success Launches By Site



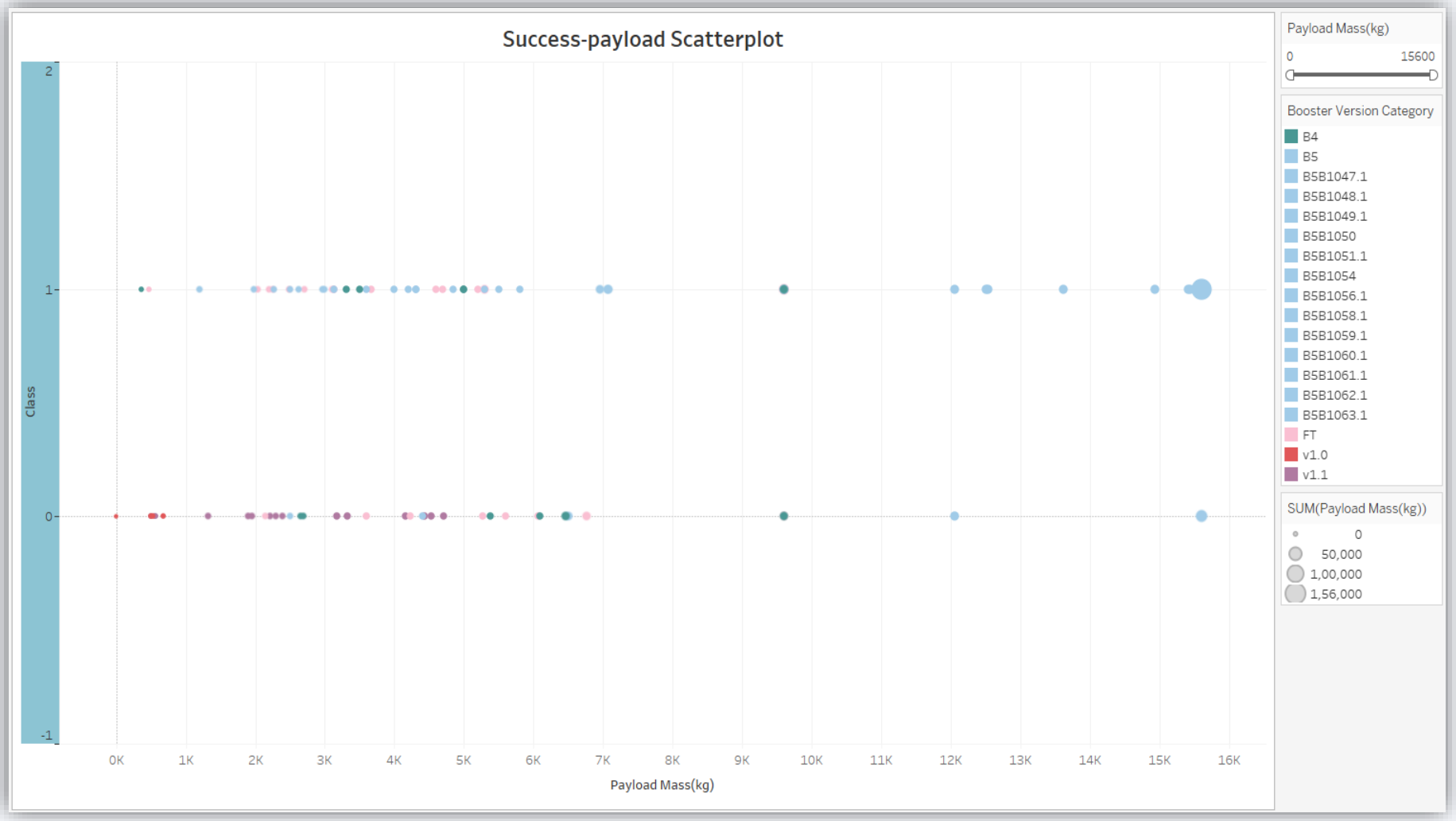
This chart shows the success rate of each site. We can clearly see that launch site CCAFS SLC-40 has the highest no of successful launches with 40.98%

Launch Success Rate



This chart shows the launch site with highest success rate. We can see that the launch site KSC LC-39A has the highest success rate of 80%

Payload vs. Launch Outcome Scatterplot



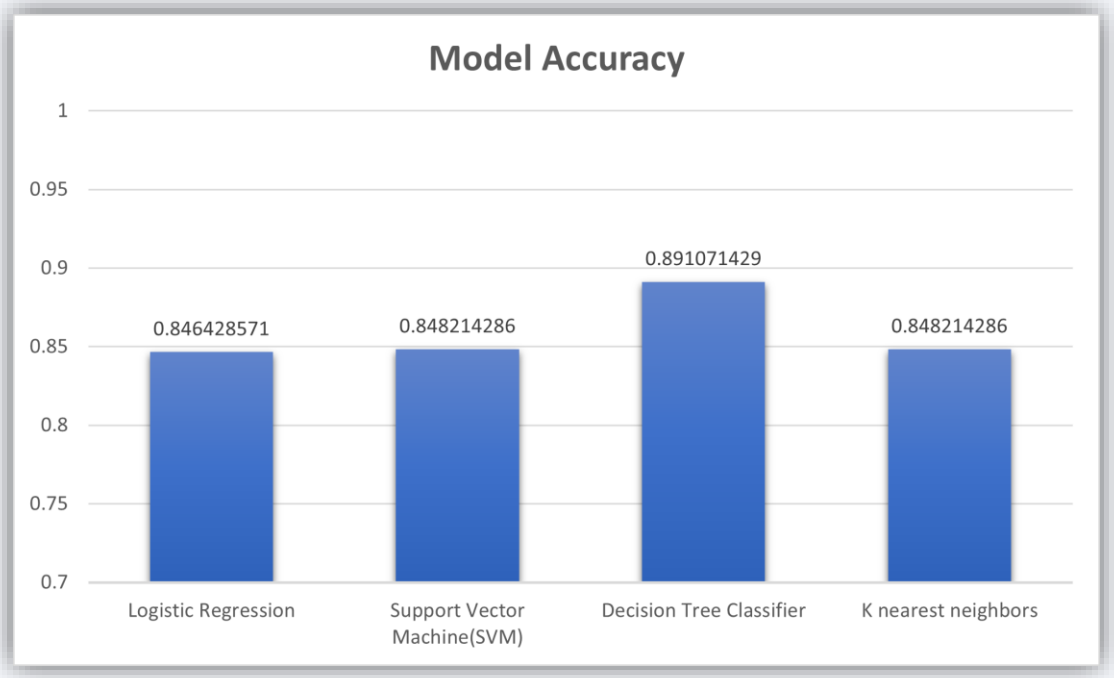
This chart shows the scatterplot between payload and Launch Outcome. There is also a slider where we can adjust the range of payload mass. The points have different colours according to booster version categories which helps us to see the launch success of each boosters.

Section 6

Predictive Analysis (Classification)

Classification Accuracy

- Bar chart of all the built classification models with built model accuracy



We can infer from the chart that Decision Tree Classifier has the built model accuracy and that we should use this model for prediction

Confusion Matrix

Confusion matrix of Decision Tree Classifier Model

We can see that out of all the predictions only 3 times the model predicted that the first stage will land whereas it actually didn't land.

Conclusions

Summary of our Analysis:

i. The success rate of launch i.e., landing of first stage successfully since 2013 kept increasing till 2020.

ii. We saw that as the Flight Number(indicating continuous launch attempts) increases, the first stage is more likely to land successfully whereas if payload increases, it is less likely to return.

iii. Orbit types ES-L1, GEO, HEO, SSO have the highest success rates.

iv. The launch sites are located relatively close to the coastline resulting in less transportation cost.

v. Out of 4 classification models we built Decision Tree had the highest accuracy which we used to predict the outcomes

vi. Launch site CCAFS SLC-40 has the highest no of successful launches with 40.98% of the total number of successful launches

vii. Different launch site have different success rate with KSC LC-39A having the highest at 80%.

viii. B5 Booster version has the highest launch success rate

Conclusions

Outcome:

With the insights we gathered from our analysis, by visualizing our dashboard and by using the model we built, we can conclude that we will be able to find the cost estimate of each launch which is exactly what was in our mind while performing this analysis.

Appendix

- GitHub Repository Link: https://github.com/devoeop/IBM_capstone
- Tableau profile Link: https://public.tableau.com/app/profile/debabrata.garai
- IBM Data Science Capstone project Link: https://www.coursera.org/learn/applied-data-science-capstone?specialization=ibm-data-science

Acknowledgements

Primary Instructors: Joseph Santarcangelo, Yan Luo

Lab Guides: Joseph Santarcangelo, Yan Luo, Azim Hirjani, Lakshmi Holla

Teaching Assistants: Malika Singla, Duvvana Mrutyunjaya Naidu

Thank you!