IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Nguyen Yen
Sept 7, 2021

Outline

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

Executive Summary

- Beautiful Soup and Python Request libraries were used for data collection
- The data was housed on an IBM Cloud DB2 instance for exploratory data analysis, and SQL was used to query the data.
- Matplotlib and seaborn were used to create the plots and charts in Python. On folium maps, the geographic data was visualized.
- Finally, we used Python's sklearn machine learning framework to create various predictive models.
- The Plotly Dash framework was also used to create an interactive dashboard.
- After evaluating and comparing the models, the model with the highest accuracy was chosen.

Introduction

- Due to the recovery of their first stage rockets, Space X is able to execute space flight at a substantially lower cost.
- Our firm, Space Y, wants to estimate if Space X will recover the first stage after a launch or not.
- If we can anticipate if the first stage will be recovered, we may estimate that the overall cost of the launch will be lower than that of competitors.
- As a result, we plan to develop a predictive model to answer the same question and assist us (Space Y) in making a more competitive offer for rocket launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Public available API endpoints from Space X
- Perform data wrangling
  - Looking specifically at Falcon 9
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Data was acquired using Space X's publicly accessible API endpoints.
- All of the launch details for Falcon rockets were included in the data.
- The requests package in Python was used to send GET requests to the endpoints and download data in JSON format, which was then transformed into a Pandas Data Frame.
- Web scraping from the wiki website yielded historical records of Falcon 9 launches.
- https://api.spacexdata.com/v4

Data Collection - SpaceX API

- The launch-related data was obtained via the Space X API.
- • The payload, cores, launchpads, and rockets were employed as endpoints.
- https://github.com/iamacsimet/Data-Science-Project-about-SpaceX/tree/main/Lab 1

Data Collection - Scraping

- We utilized Beautiful soup to web scrape data from html pages into a pandas data frame, and we used Falcon 9 historical launch data from Wikipedia.
- https://github.com/iamacsimet/Data-Science-Project-about-SpaceX/tree/main/Web Scrapping Lab

Data Wrangling

- We filtered away rows from the primary data frame that only contained Falcon 9 information.
- The Payload Mass column had some missing values, which were eliminated and replaced by the column's overall mean.

- You need to present your data wrangling process using key phrases and flowcharts
- https://github.com/iamacsimet/Data-Science-Project-about-SpaceX/tree/main/Lab 1

EDA with Data Visualization

- Several plots were created using Python packages such as matplotlib and seaborn.
- The plots were created to visualise and comprehend the many relationships that the data features had with one another; for example, a cat plot was created for the “Flight No. vs. Launch Site” relationship; we also plotted scatter plots, bar charts, and line graphs for the same purpose.
- https://github.com/iamacsimet/Data-Science-Project-about-SpaceX/tree/main/Lab 4

EDA with SQL

- SQL is a language that may be used to query and manipulate data in a database.
- Data from Space X launches was saved on IBM Cloud in a DB2 instance.
- We were able to connect to the IBM Cloud hosted database and run SQL queries from Python using sqlalchemy, ibm db sa, and ipython-sql.
- https://github.com/iamacsimet/Data-Science-Project-about-SpaceX/tree/main/Lab 3

Build an Interactive Map with Folium

- To examine data on the map views, Folium, a python-based framework for creating map visualisations, was utilised.
- Because a successful launch is dependent on the position of the launch site and its closeness to other sites such as the sea shore, townships, rivers, and other landmarks, it is critical to select the best launch location. Folium was used to create interactive maps for this reason.
- https://github.com/iamacsimet/Data-Science-Project-about-SpaceX/tree/main/Lab 5

Build a Dashboard with Plotly Dash

- Plotly Dash was used to create a small web app (based on the Flask micro framework).
- The web app is a straightforward dashboard with two graphs.

  - Pie Chart – Success Rate vs Launch Sites
  - Scatter Chart – Launch Outcome vs Payload Mass.

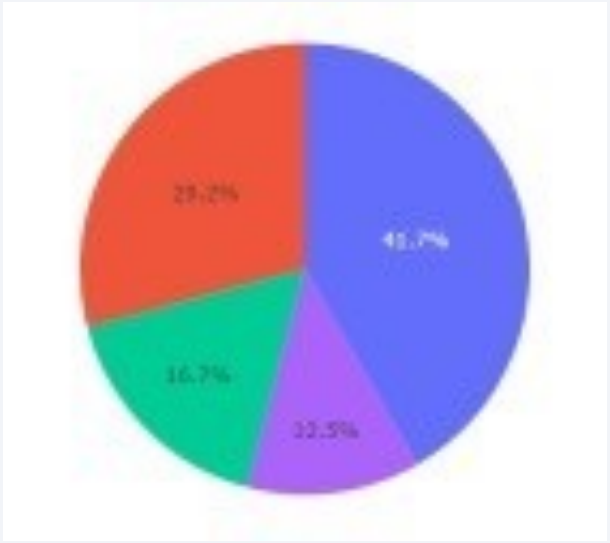
- We also created a dropdown menu to allow the user to choose which launch location they want to see the statistics for. In the second scatter figure, we also incorporated a slider to select the payload mass range.
- https://github.com/iamacsimet/Data-Science-Project-about-SpaceX/blob/f161a4c117d86af21fce2444d2696f22ec94ea9a/application-dash.py

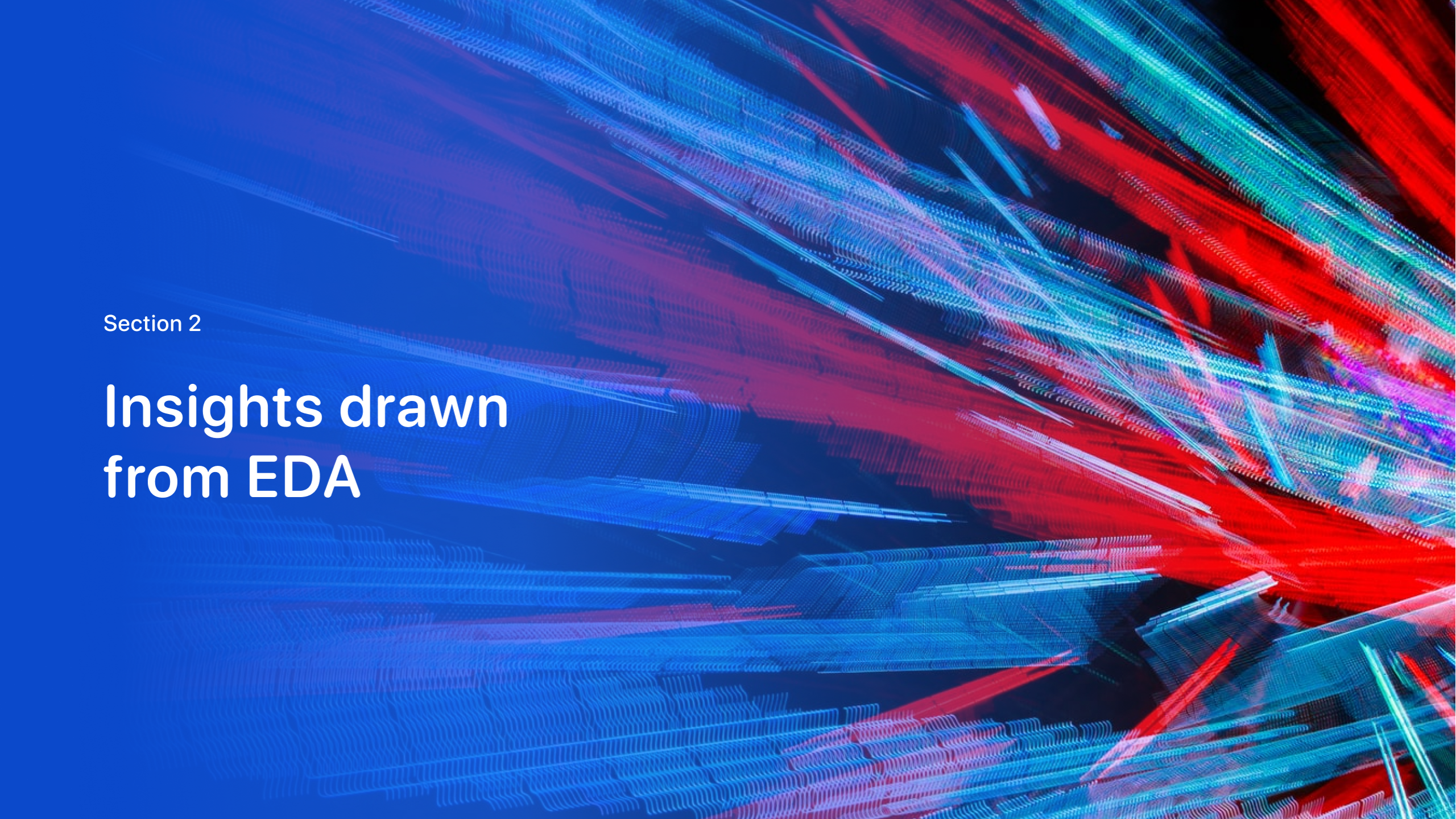
Predictive Analysis (Classification)

- Based on available input features such as Payload Mass, Flight No, Orbit Type, and so on, predictive models such as logistic regression, KNN, SVM, and Decision Trees were used to create predictions on whether the launch would be successful or not.
- The input data was divided into two sets: training and test. The training set was used to create the models, while the test sets were used to assess the models' accuracy.
- We determined which model worked better based on the accuracy results. The Grid Search method was also used to determine the hyper parameters of each model.
- https://github.com/iamacsimet/Data-Science-Project-about-SpaceX/tree/main/Lab 6

Results

- A variety of data plots and map visualizations were provided in our EDA.
- We were able to see connections between distinct input features and the consequences that resulted from them.
- Scatter plots such as Flight No. vs. Outcome revealed that the outcome of a launch was largely influenced by payload tonnage and the orbit chosen for the launch.
- Using Folium, we visualized where successful launches occurred and where unsuccessful launches occurred. We also had to calculate the distances between the launch zones and nearby town ships or coasts.
- Interactive analytics demo in screenshots
- Predictive analysis results

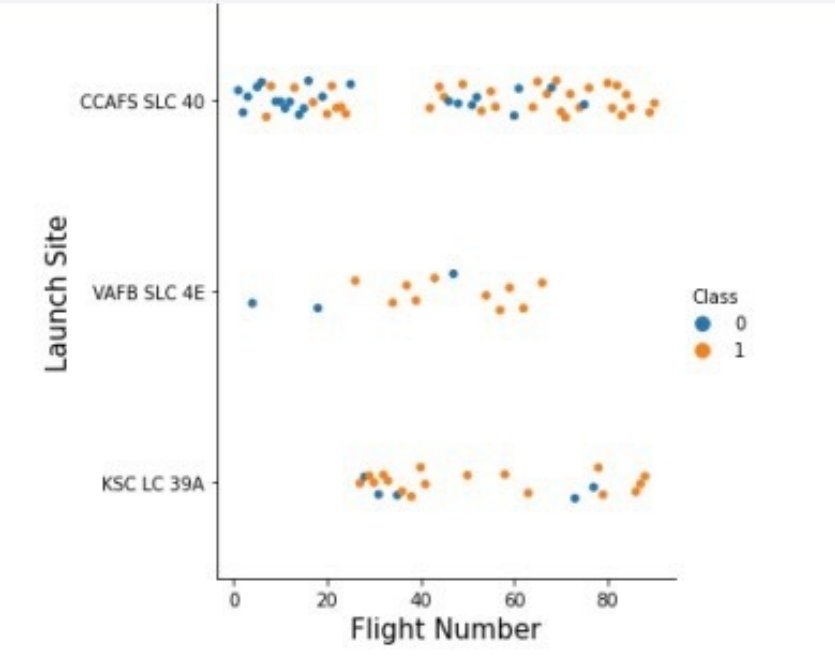




Section 2

Insights drawn from EDA

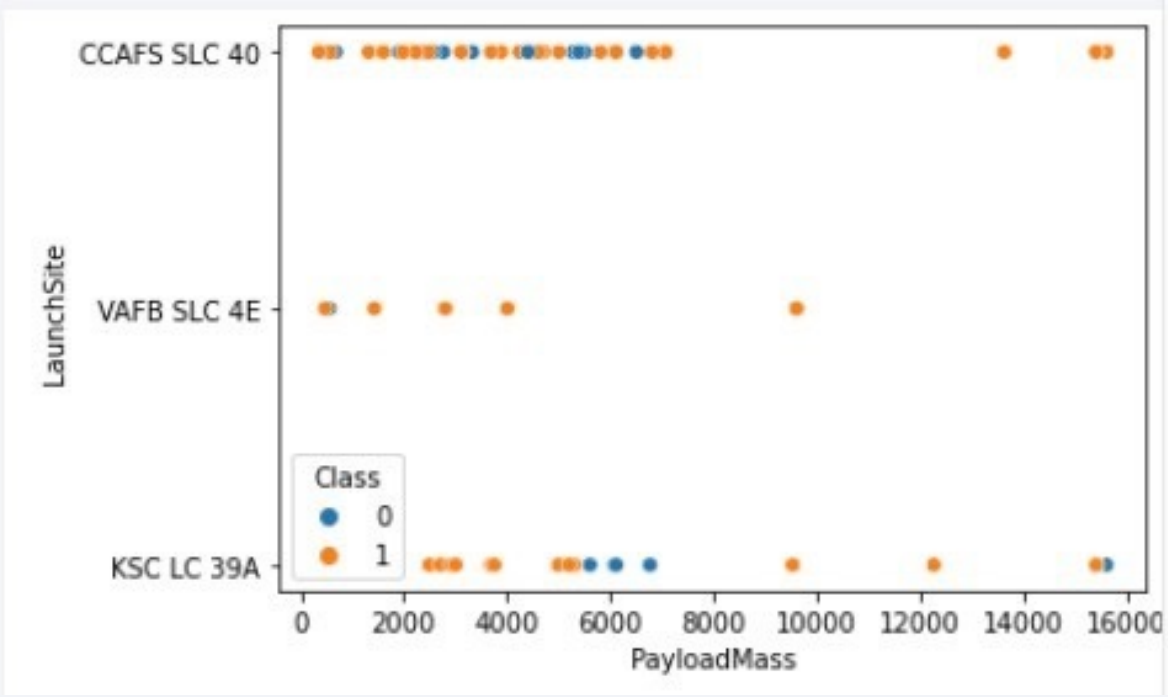
Flight Number vs. Launch Site



- According to the graph, the launch site CCAFS SLC 40 has more launches than any other launch site; however, the flight number does not appear to have a clear association with the flight's outcome. The same can be stated for the launch site, but when compared to CCAFS SLC 40, we can observe that KSC LC 39 A and VAFB SLC 4E have a lower failure rate.

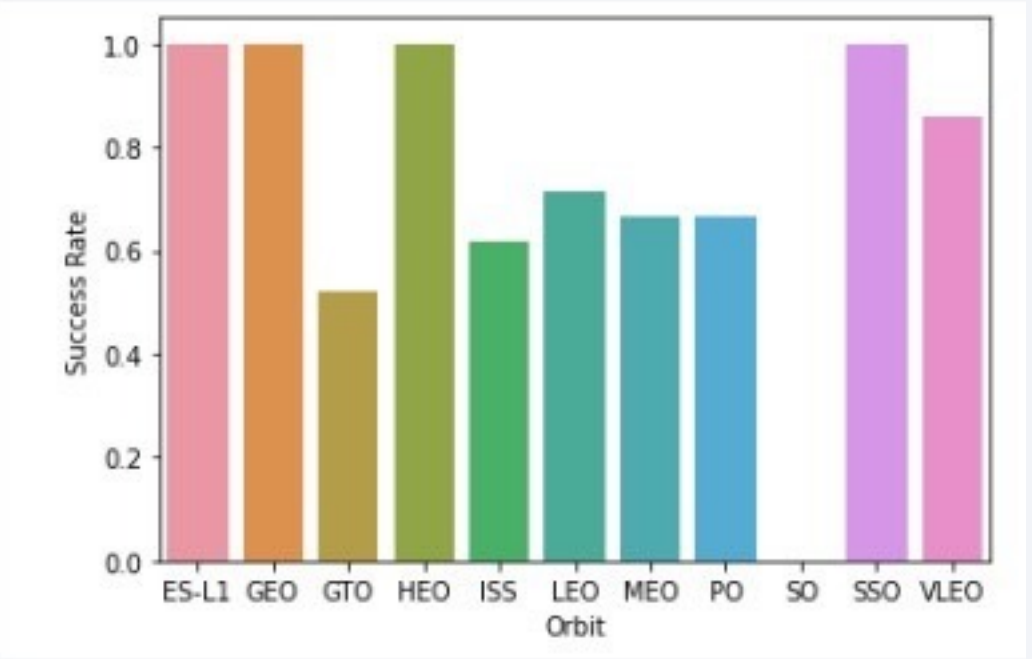
Payload vs. Launch Site

- The payload varies from 0 kg to16000 kg
- From KSC LC 39, the launches must have a payload of at least 2000 kilos
- Launches with a larger payload greater than 8000 kg appears to have a positive consequence
- CCAFS SLC 40 is the one with the most Payloads vary depending on the launch. between 1000 and ten thousand kilograms



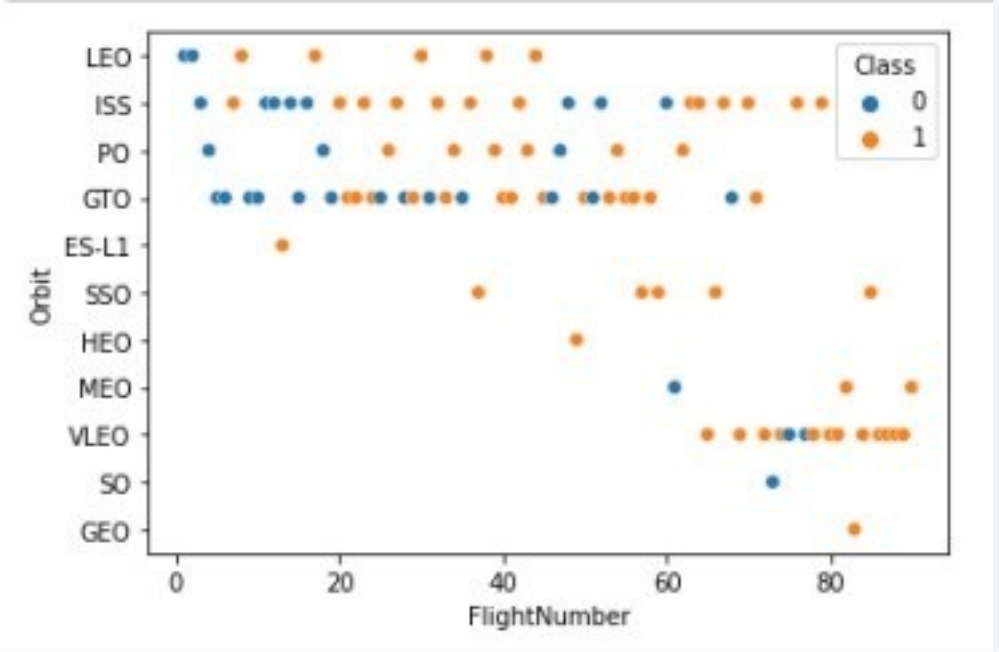
Success Rate vs. Orbit Type

- The Orbit and the Success Rate
- • As can be seen, GTO has the lowest success rate, while ES-L1, GEO, HEO, and SSO have the best success rates.



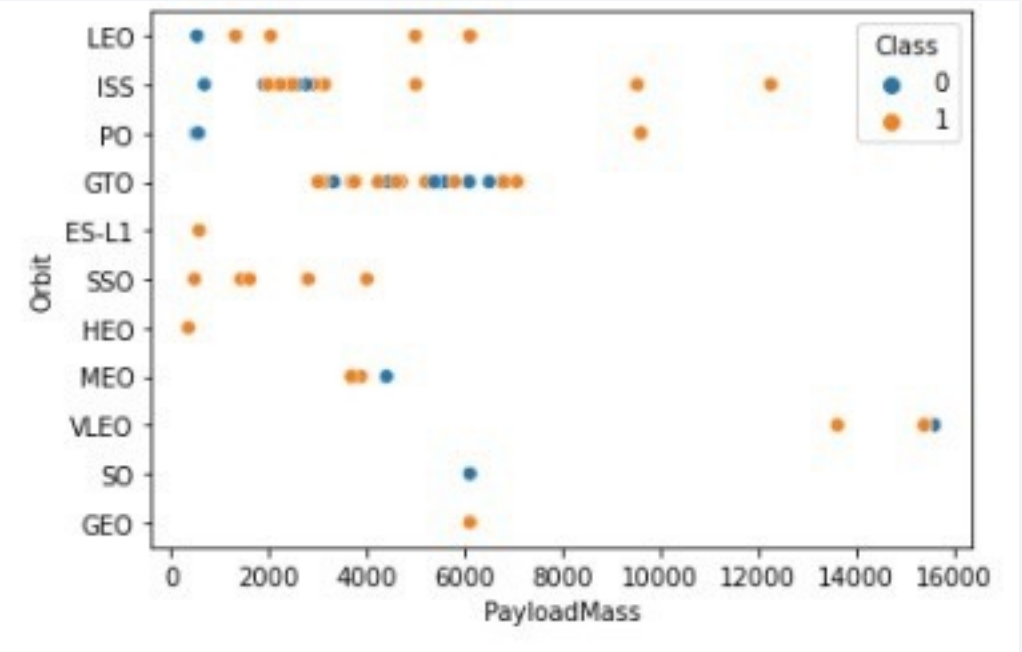
Flight Number vs. Orbit Type

- Scatter plot for Orbit Type vs. Flight.
- In LEO orbit, success appears to be connected to the number of flights; however, in GTO orbit, there appears to be no relationship between flight number and success.



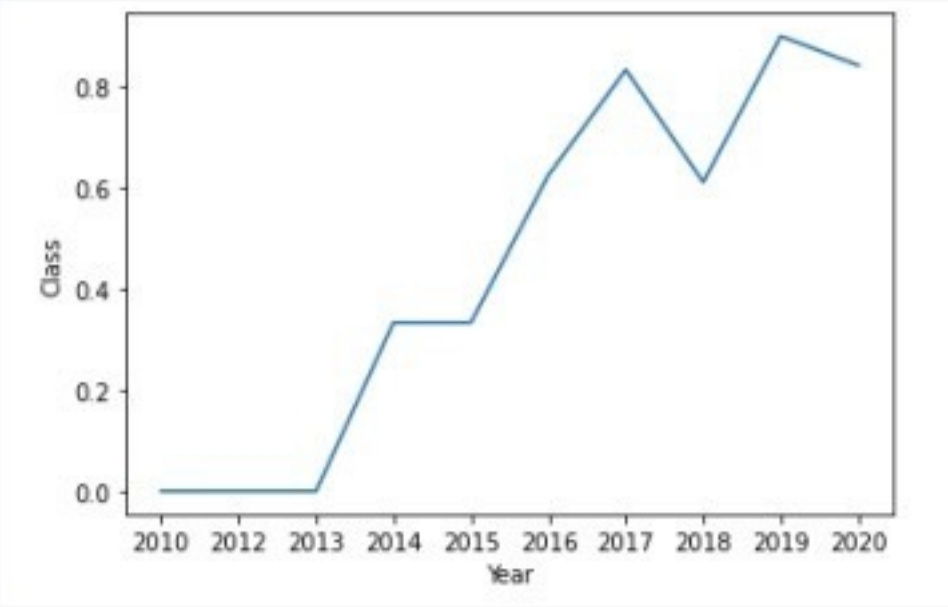
Payload vs. Orbit Type

- Payload Mass vs Orbit type scatter plot
- Heavy payloads have a detrimental impact on GTO orbits but a favourable impact on GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend

- Trend for success vs year
- Rate has significantly increased post 2013 to 2020



All Launch Site Names

- Unique launch sites:

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

- Query:

```
select Unique(LAUNCH_SITE) from
SPACEXTBL;
```

Launch Site Names Begin with 'CCA'

- Query:

```
select * from SPACEXTBL where LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

- Results:

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	None	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	None	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	None	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	None	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	None	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- Total payload mass: 45,596 KG

- Query:

```
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE
CUSTOMER='NASA (CRS)‘
```

Average Payload Mass by F9 v1.1

- Result: 2534 KG

- Query Used

```
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE
BOOSTER_VERSION LIKE 'F9 v1.1%';
```

First Successful Ground Landing Date

- Result: December 22, 2015

- Query Used

```
SELECT MIN(DATE) FROM SPACEXTBL WHERE MISSION_OUTCOME='Success'
AND LANDING__OUTCOME='Success (ground pad)';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

- Query Used: (results to the right)

```
select distinct(booster_version) from SPACEXTBL where
landing__outcome='Success (drone ship)' and
PAYLOAD_MASS__KG_>=4000 and PAYLOAD_MASS__KG_<=6000;
```

booster_version
F9 B4 B1041.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1031.2
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1036.1
F9 FT B1038.1

Total Number of Successful and Failure Mission Outcomes

- Successful and failure outcomes:

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

- Query Used:

```
select mission_outcome, count(mission_outcome) as count from SPACEXTBL group by
mission_outcome;
```

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Query Used:

```
select distinct(booster_version) from SPACEXTBL
where payload_mass__kg_=(select
max(payload_mass__kg_) from SPACEXTBL);
```

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- List the failed landingoutcomes in drone ship, their booster versions, and launch site names for in year 2015
- Query Used:

```
select booster_version, launch_site from SPACEXTBL where
landing__outcome='Failure (drone ship)' and YEAR(date)='2015';
```

- Results

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
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

- Query Used:

```
select count(landing__outcome) as count, landing__outcome from
SPACEXTBL where date>='2010-06-04' and date<='2017-03-20'
group by landing__outcome order by count(landing__outcome) desc;
```

- Results on right

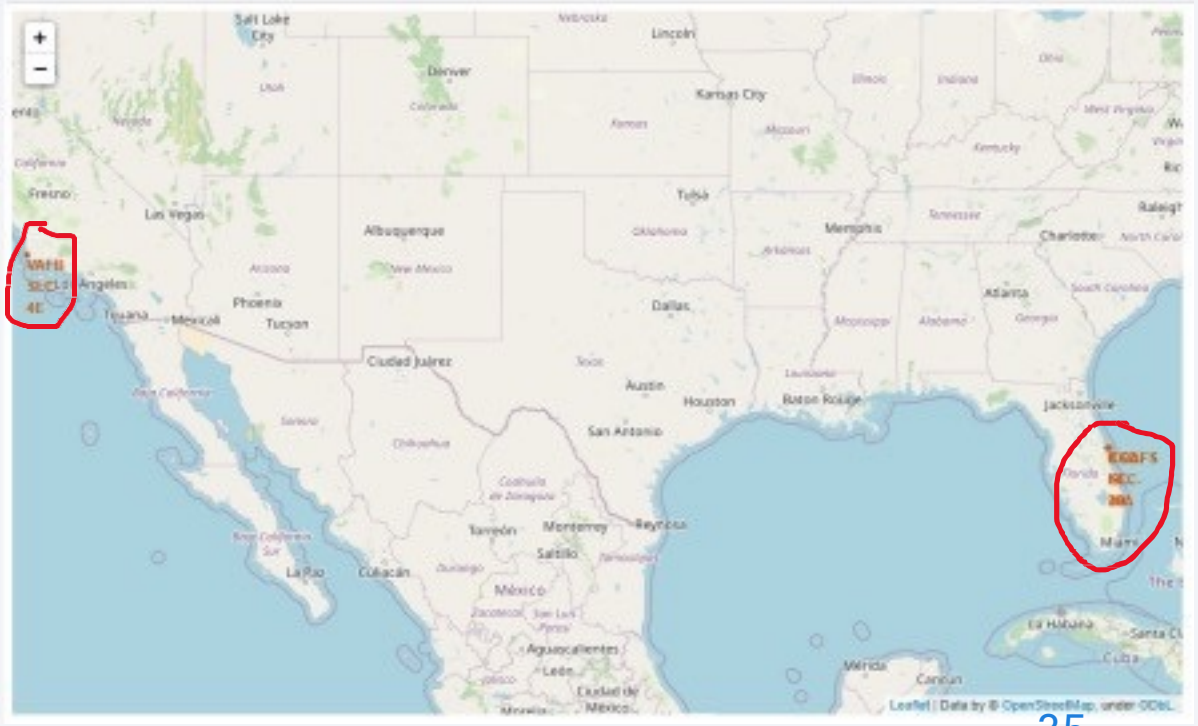
COUNT	landing__outcome
10	No attempt
5	Failure (drone ship)
5	Success (drone ship)
3	Controlled (ocean)
3	Success (ground pad)
2	Failure (parachute)
2	Uncontrolled (ocean)
1	Precluded (drone ship)

Section 4

Launch Sites Proximities Analysis

Site for Launch Locations

- The following map depicts the many launch sites as well as the NASA JSC.
- They're all marked on a typical map of the United States.
- Majority of the launch points are along the water's edge - One is in the Los Angeles area, and the others are in Florida.

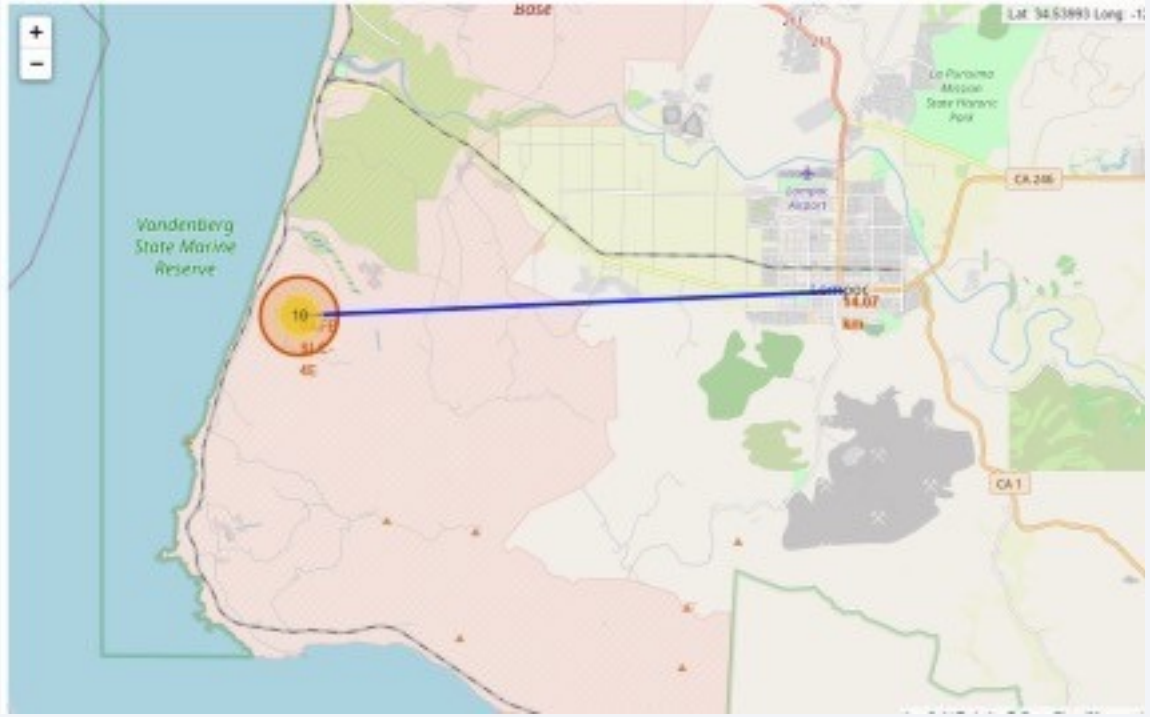


Success and Failures for launches at sites

- We put markers to the map below to show the launch results for each launch point.

- We were able to identify and pinpoint which launch sites had the most successful launches using the map.

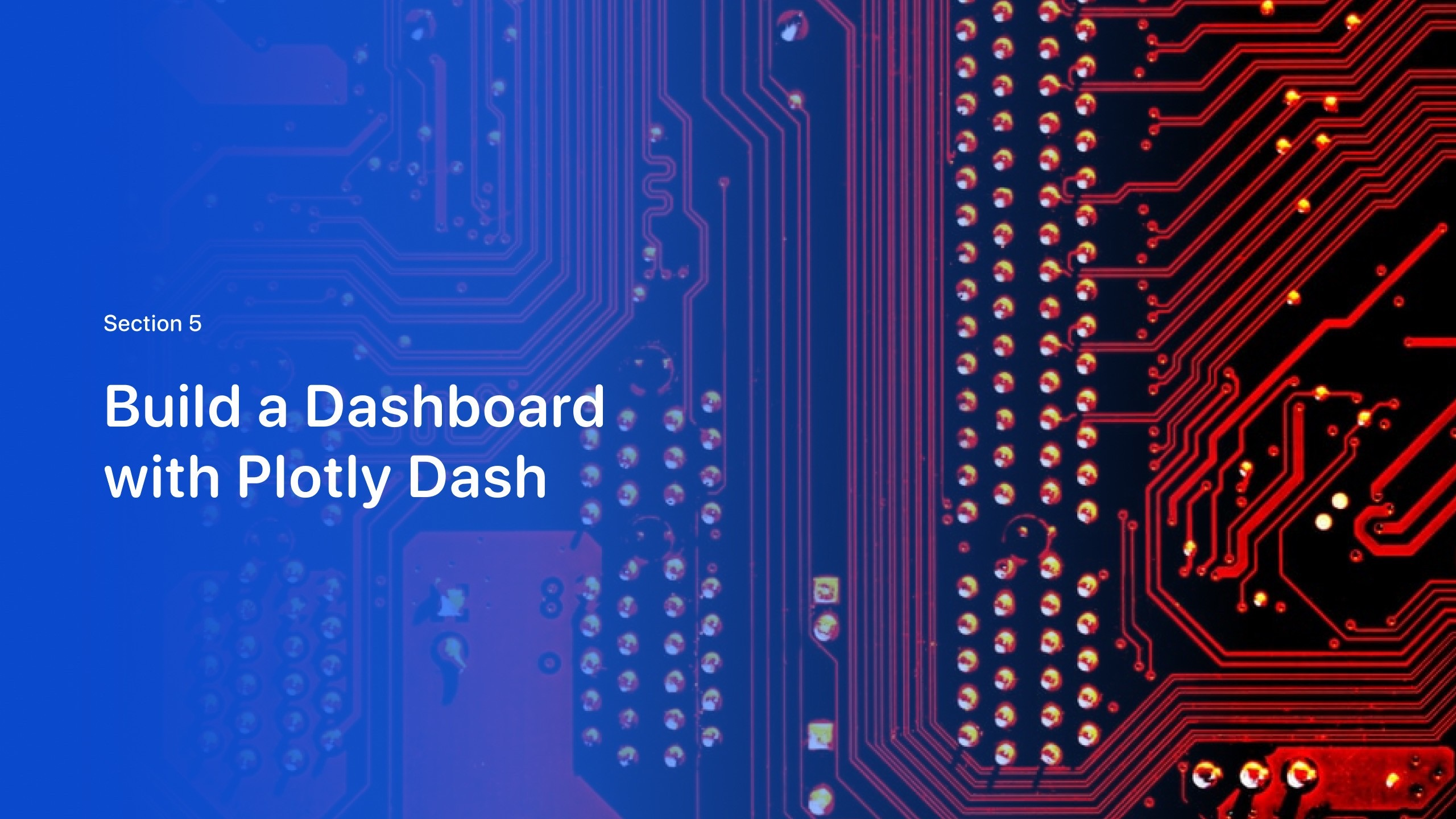
Launch Site Distances



- We can also create polygons, lines, and other geometries on the map using folium.

•Calculated the distance between two points and then drew a line between them.

•As can be seen on the map below, there is an airport (Lompoc Airport) within 14.07 KM of the launch location - VAFE SLC 4E.


Section 5
Build a Dashboard with Plotly Dash

Launch Sites Success Rates

- Shows success percent for launch sites
- KSC LC 39-A has highest success rate
- CCAFS SLC 40 has lowest success rate

In Depth: KSC LC 39-A

- Chart for KSC LC 39-A through Plotly dashboard
- Success rate of 77%
- Failure rate of 23%

Launch vs Payload: Outcomes

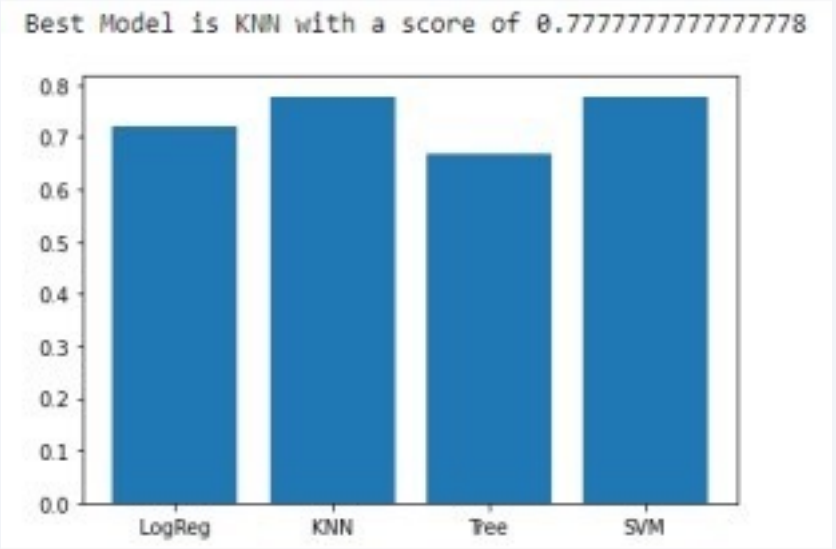
- The link between launch outcomes and payload mass is depicted in the scatter plot below. We have a payload mass range slider that allows us to filter a payload mass range for the scatter plot. We've chosen a weight range of 2000kg to 6000kg.
- The points are color-coded according to the booster version that was utilised.
- We can observe that many of the v1.1 boosters failed to launch, but FT boosters appear to be more successful than the rest.

Section 6

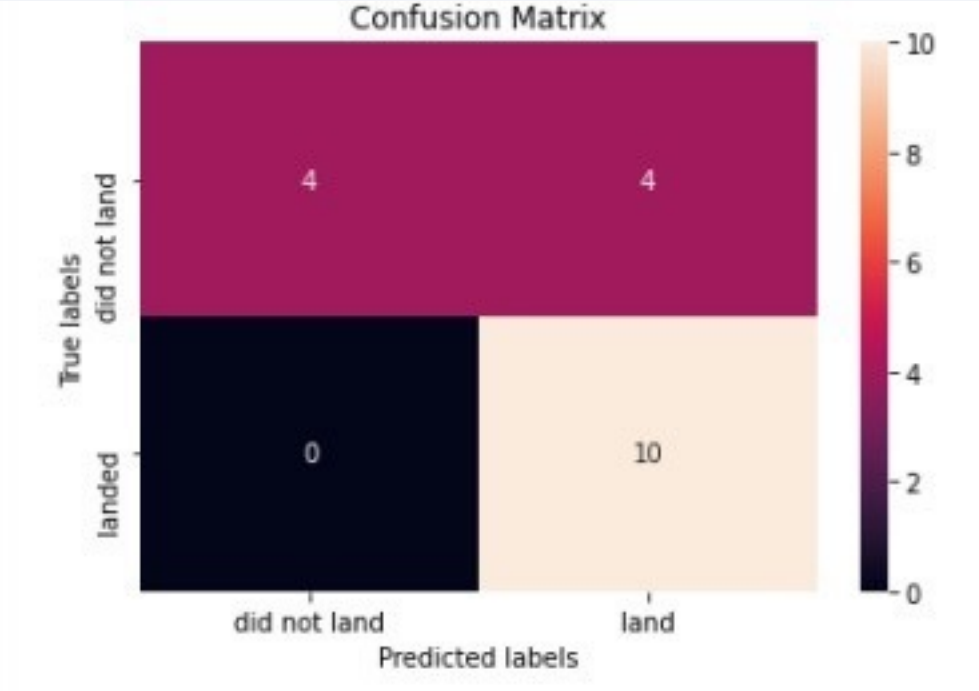
Predictive Analysis (Classification)

Classification Accuracy

- Models were trained and tested on same dataset
- KNN performed best
- Max Accuracy we could get was 77%



Confusion Matrix



- Confusion matrix made for KNN model
- Model could predict all 10 correctly of the 10 flights that were successful
- Of the 8 unsuccessful flights, 4 were predicted correctly

Conclusions

- With the help of many input data and attributes provided by Space X Falcon 9 flights, we attempted to construct a prediction model.
- The KNN models were chosen as the best model for the given data set and input attributes after a comparison of other machine learning models.
- We may be able to enhance the model's accuracy in the future by giving more data or using other machine learning approaches such as neural networks; nonetheless, we were able to evaluate and understand the data and display it in a well-structured manner using dashboards and charts.
- Before moving on to creating predictive models, these are required for any data science endeavor.

Thank you!