Winning Space Race with Data Science

Shilpa Shinde

12 August 2022

Outline

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

Executive Summary

- **Summary of methodologies**
  - Data collection
  - Data wrangling
  - EDA with data visualization
  - EDA with SQL
  - Building an interactive map with Folium
  - Building a Dashboard with Plotly Dash
  - Predictive analysis (Classification)
- **Summary of all results**
  - Exploratory data analysis results
  - Interactive analytics demo
  - Predictive analysis results

Introduction

- **Project background and context**

The era of commercial space has arrived, and there are several companies that are making space travel affordable for everyone. Perhaps the most successful of them is SpaceX, and one o the reasons is that their rocket launch is relatively inexpensive.

SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage.

Therefore, we will predict if the Falcon 9 first stage will land successfully. If we can determine if the first stage will land, we can determine the cost of a launch. This information can be used an alternate company wants to bid against SpaceX for a rocket launch.

- **Problems you want to find answers**

- Correlations between each rocket variables and successful landing rate
- Conditions to get the best results and ensure the best successful landing rate

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API & Web Scraping Falcon 9 and Falcon Heavy Launches Record from Wikipedia
- Perform data wrangling
  - Convert outcomes into Training Labels with the booster successfully/unsuccessful landed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Das
- Perform predictive analysis using classification models
  - Find best Hyperparameter for Logistic Regression, SVM, Decision Trees

Data Collection

The data collection process includes a combination of API requests from the

SpaceX API and web scraping data from a table in the Wikipedia page of

SpaceX, Falcon 9 and Falcon Heavy Launches Records.

- SpaceX API Data Columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSi
  Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude
- Wikipedia Web Scrape Data Columns: Flight No., Launch site, Payload, PayloadMass, Orbit, Cus
  Launch outcome, Version Booster,Booster landing, Date, Time

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL of the completed SpaceX API calls notebook (must include completed code cell and outcome cell), as an external reference and peer-review purpose

```
# Takes the dataset and uses the rocket column to call the API and append the data to the list
def getBoosterVersion(data):
    for x in data['rocket']:
        response = requests.get("https://api.spacexdata.com/v4/rockets/"+str(x)).json()
        BoosterVersion.append(response['name'])
```

From the `launchpad` we would like to know the name of the launch site being used, the logitude, and the latitude.

```
# Takes the dataset and uses the launchpad column to call the API and append the data to the list
def getLaunchSite(data):
    for x in data['launchpad']:
        response = requests.get("https://api.spacexdata.com/v4/launchpads/"+str(x)).json()
        Longitude.append(response['longitude'])
        Latitude.append(response['latitude'])
        LaunchSite.append(response['name'])
```

From the `payload` we would like to learn the mass of the payload and the orbit that it is going to.

```
# Takes the dataset and uses the payloads column to call the API and append the data to the lists
def getPayloadData(data):
    for load in data['payloads']:
        response = requests.get("https://api.spacexdata.com/v4/payloads/"+load).json()
        PayloadMass.append(response['mass_kg'])
        Orbit.append(response['orbit'])
```

Data Collection - Scraping

1.Getting response from HTML

2.Creating a BeautifulSoup object

3.Finding all tables and assigning the result to a list

4.Extracting column name one by one

5.Creating an empty dictionary with keys

6.Filling up the launch_dict with launch records

7.Creating a Dataframe and exporting it to a CSV

```
html_data = requests.get(static_u

soup = BeautifulSoup(html_data, '

html_tables = soup.find_all(

df=pd.DataFrame(launch_
```

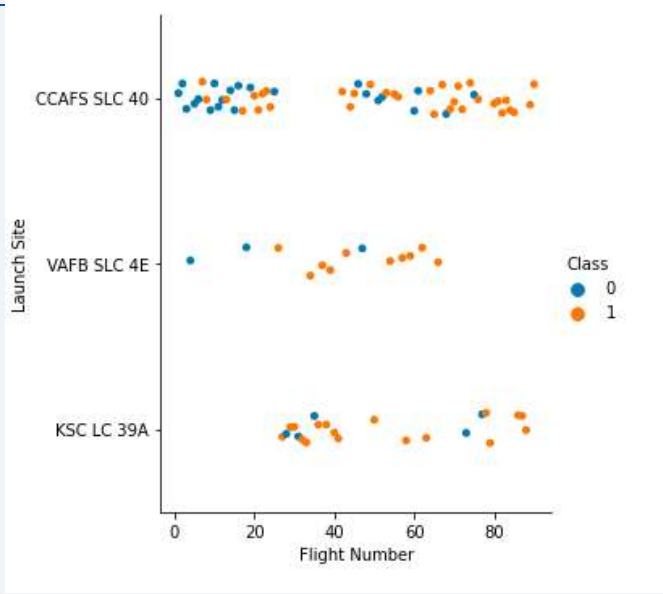
Data Wrangling

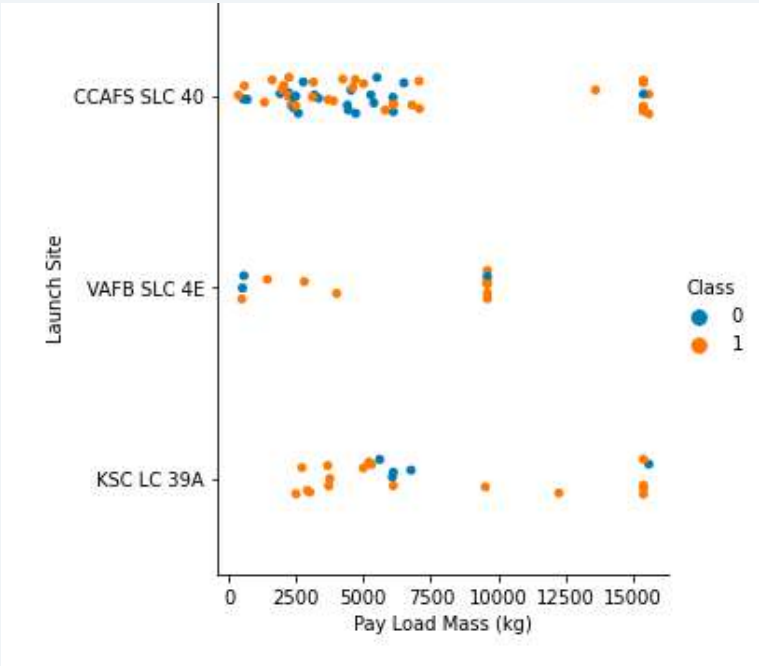
Describe how data were processed
You need to present your data wrangling process using key phrases and flowcharts
Add the GitHub URL of your completed data wrangling related notebooks, as an external reference and peer-review purpose

EDA with Data Visualization

Relationship between

Flight Number and Launch Site

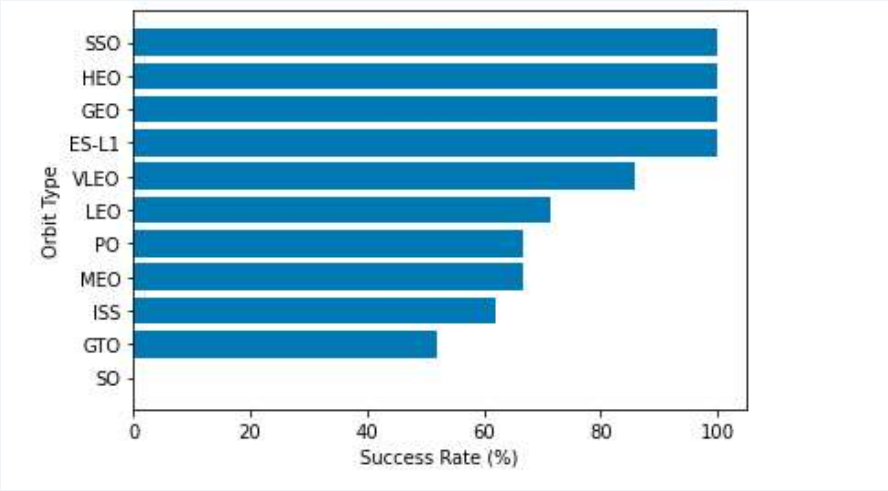


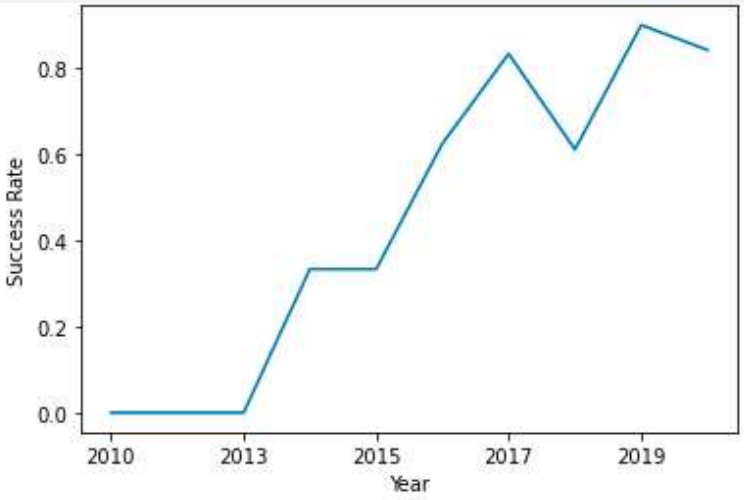


Relationship between Payload and Launch S

EDA with Data Visualization

Relationship between success rate of each orbit type





you can observe that the sucess rate since 2013 kept increasing till 2020

Launch success yearly trend

EDA with SQL

- %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
- %sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
- %sql SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass_kg FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
- %sql SELECT AVG(PAYLOAD_MASS__KG_) AS avg_payload_mass_kg FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1'

Build an Interactive Map with Folium

- Objects created and added to a folium map:
  - Markers that show all launch sites on a map
  - Markers that show the success/failed launches for each site on the map
  - Lines that show the distances between a launch site to its proximities
- By adding these objects, following geographical patterns about launch sites are found:
  - Are launch sites in close proximity to railways? Yes
  - Are launch sites in close proximity to highways? Yes
  - Are launch sites in close proximity to coastline? Yes
  - Do launch sites keep certain distance away from cities? Yes

Build a Dashboard with Plotly Dash

- The dashboard application contains a pie chart and a scatter point chart
  - Pie chart
    - For showing total success launches by sites
    - This chart can be selected to indicate a successful landing distribution across all launch sites or to indicate the success rate of individual launch sites.
  - Scatter chart
    - For showing the relationship between Outcomes and Payload mass(Kg) b different boosters
    - Has 2 inputs: All sites/individual site & Payload mass on a slider between and 10000 Kg
    - This chart helps determine how success depends on the launch point, payload mass, and booster version categories.

Predictive Analysis (Classification)

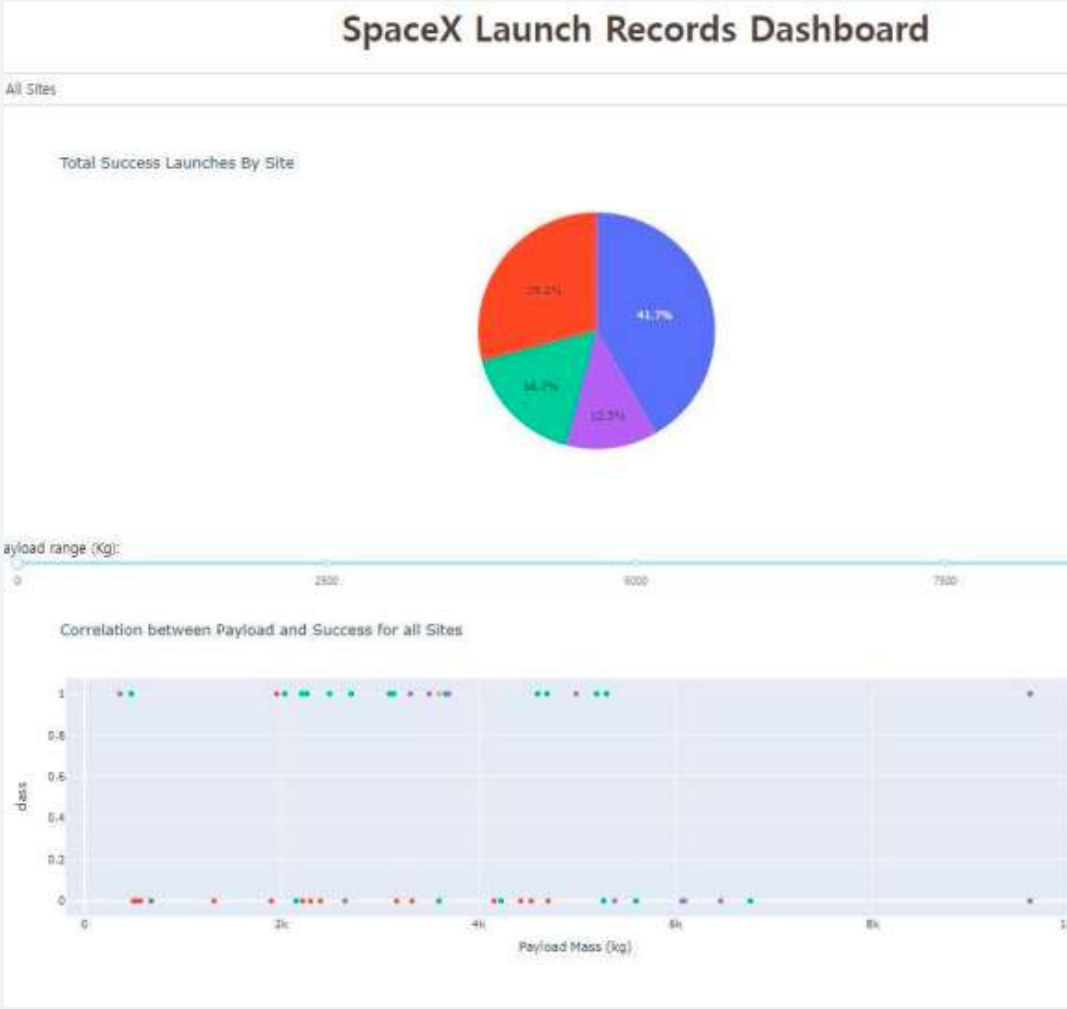
- Perform exploratory Data Analysis and determine Training Labels
  - Create a column for the class
  - Standardize the data
  - Split into training data and test data
- Find best Hyperparameter for SVM, Classification Trees and Logistic Regression
  - Find the method performs best using test data

Results

- The left screenshot is a preview of the Dashboard with Plotly Dash.

- The results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and Interactive Dashboard
will be shown in the next slides.

- Comparing the accuracy of the four methods, all return the same accuracy of about 83% for test data.

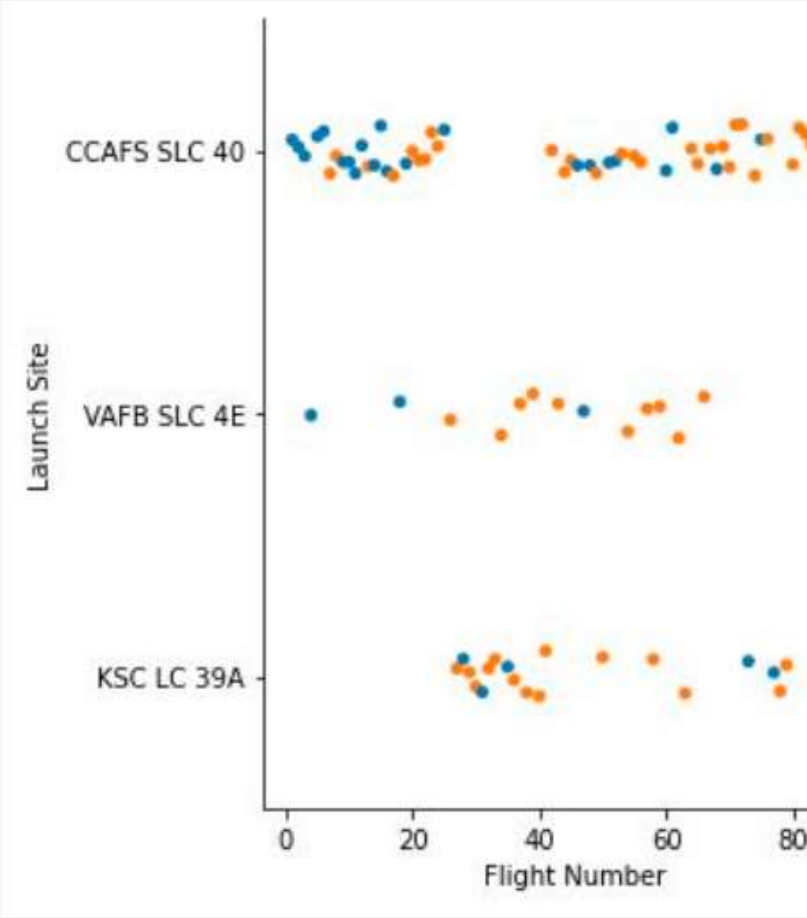


Section 2

Insights drawn from EDA

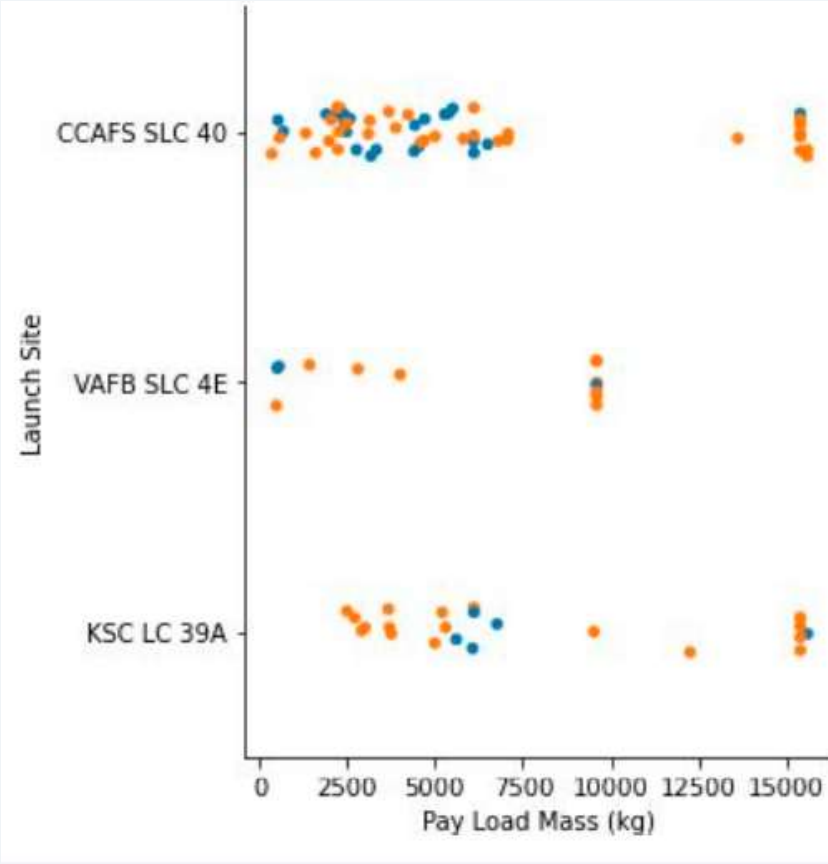
Flight Number vs. Launch Site

- This figure shows that the success rate increased as the number of flights increased.
- As the success rate has increased considerably since the 20th flight, this point seems to be a big breakthrough.



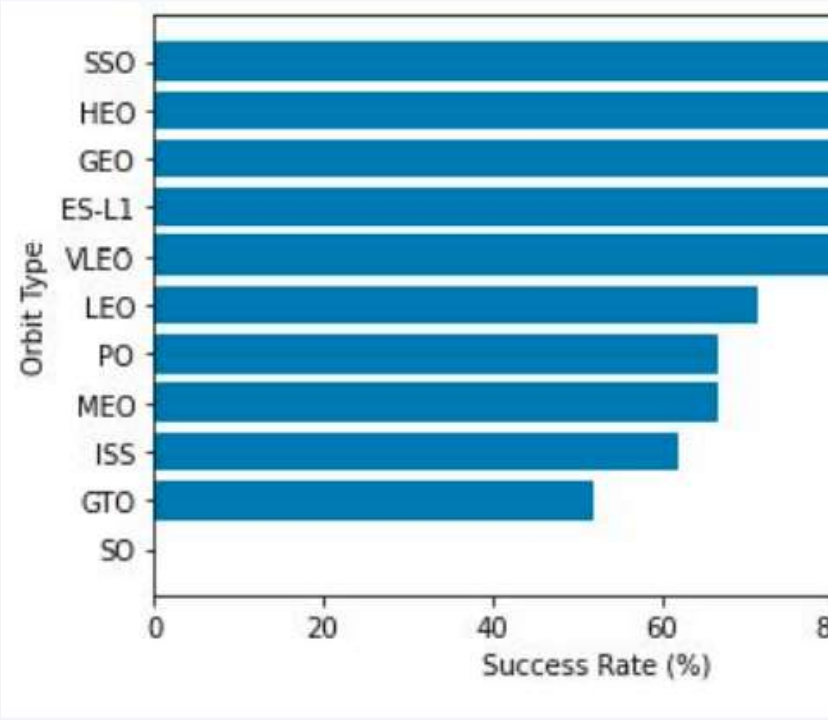
Payload vs. Launch Site

- Class 0 (blue) represents unsuccessful launch, and Class 1 (orange) represents successful launch
- At first glance, the large payload mass, the higher rocket's success rate, but it is difficult to make decisions based on this figure
- No relation is seem in these parameters



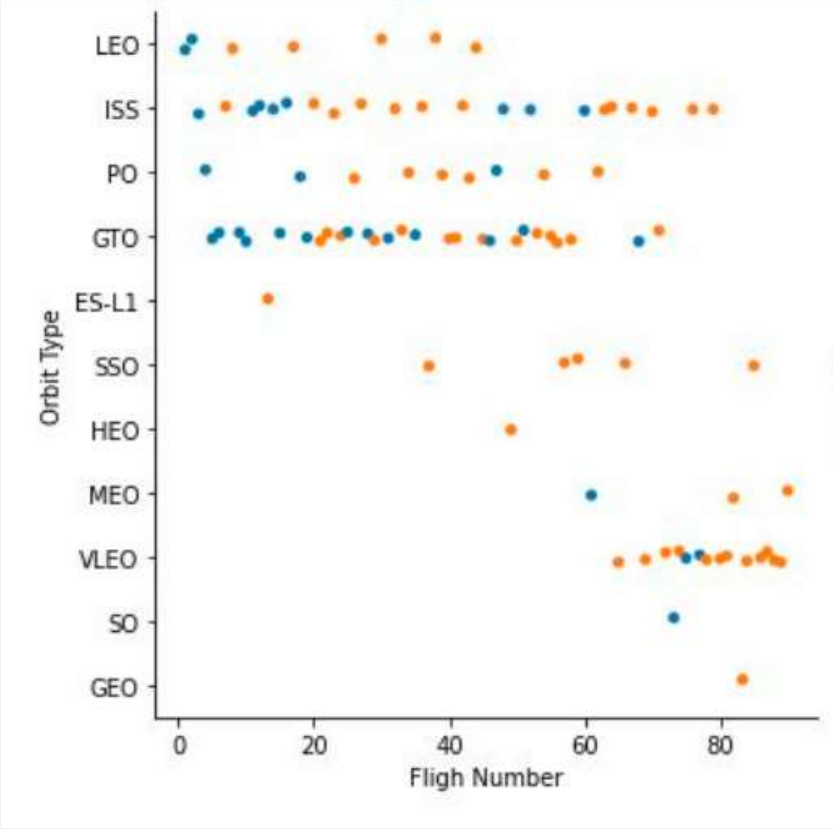
Success Rate vs. Orbit Type

- Orbit types SSO, HEO, GEO, and ES-L1 have the highest success rates (100%).
- On the other hand, the success rate of orbit type GTO is only 50%, and it is the lowest except for type SO, which recorded failure in a single attempt



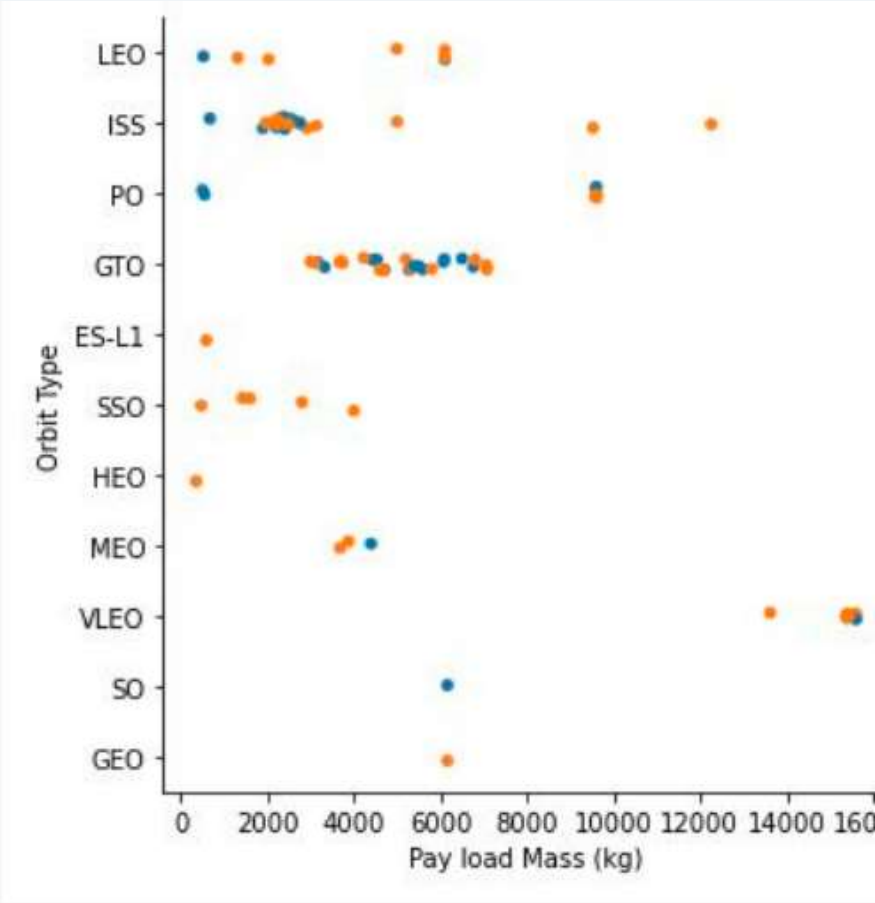
Flight Number vs. Orbit Type

- Class 0 (blue) represents unsuccessful launch, and Class 1(orange) represents successful launch.
- In most cases, the launch outcome seems to be correlated with the flight number.



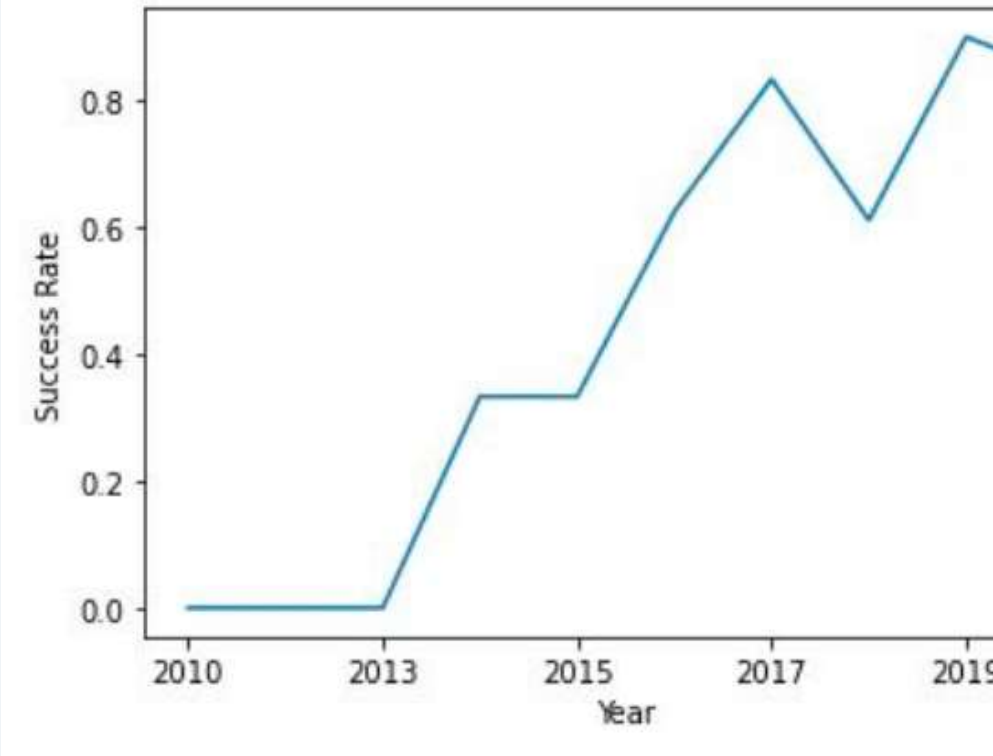
Payload vs. Orbit Type

- Class 0 (blue) represents unsuccessful launch, and Class 1 (orange) represents successful launch.
- With heavy payloads the successful landing or positive landing rate are more for LEO and ISS.



Launch Success Yearly Trend

- Since 2013, the success rate has continued to increase until 2017.
- The rate decreased slightly in 2018.
- Recently, it has shown a success rate of about 80%.



All Launch Site Names

- When the SQL DISTINCT clause is used in the query, only unique values are displayed in the Launch_Site column from the SpaceX table.
- There are four unique launch sites:

CCAFS LC-40, CCAFS SLC-40,

KSC LC-39A, VAFB SLC-4E

- Query

```
SELECT DISTINCT LAUNCH_SIT
FROM SPACEXTBL
```

- Result

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

Launch Site Names Begin with 'CCA'

- Only five records of the SpaceX table were displayed using LIMIT 5

clause in the query.

- Using the LIKE operator and the percent sign (%) together, the Launch_Site name starting with CAA could be called

```
SELECT * FROM SPACEXTBL
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure
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
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

- Using the SUM() function to calculate the sum of column PAYLOAD_MASS__KG_.
- In the WHERE clause, filter the dataset to perform calculations only if Customer is NASA (CRS)

- Query

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

- Result

total_payload_mass_kg
45596

Average Payload Mass by F9 v1.1

- Using the AVG() function to calculate the average value of column PAYLOAD_MASS__KG_.
- In the WHERE clause, filter the dataset to perform calculations only if Booster_version is F9 v1.1

- Query

```
SELECT AVG(PAYLOAD_MASS__KG_
       AS avg_payload_mass_k
FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9
```

- Result

avg_payload_mass_kg
2928

First Successful Ground Landing Date

•Using the MIN() function to find out

the earliest date in the column

DATE.

• In the WHERE clause, filter the

dataset to perform a search only if

Landing__outcome is Success

(ground pad)

• Query

```
SELECT MIN(DATE)
  AS first_successful_landing_
FROM SPACEXTBL
WHERE LANDING__OUTCOME
        = 'Success (ground pad)'
```

• Result

first_successful_landing_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- In the WHERE clause, filter the dataset to perform a search if Landing__outcome is Success (drone ship).
- Using the AND operator to display a record if additional conditio PAYLOAD_MASS__KG_ is between 4000 and 6000

- Result

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

- Query

```
SELECT BOOSTER_VERSION
FROM SPACEXTBL
WHERE LANDING__OUTCOME = 'Success (drone ship
    AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 60
```

2015 Launch Records

In the WHERE clause, filter the dataset to perform a search if Landing__outcome is
Failure (drone ship).
○ Using the AND operator to display a record if additional condition YEAR is 2015.
• In 2015, there were two landing failures on drone ships

- Query

```
SELECT LANDING__OUTCOME,
       BOOSTER_VERSION,
       LAUNCH_SITE
FROM SPACEXTBL
WHERE LANDING__OUTCOME
          = 'Failure (drone ship)'
       AND YEAR(DATE) = '2015'
```

- Result

landing__outcome	booster_version	launc
Failure (drone ship)	F9 v1.1 B1012	CCAFS
Failure (drone ship)	F9 v1.1 B1015	CCAFS

Rank Landing Outcomes Between 2010-06-04 and 2017

In the WHERE clause, filter the dataset to perform a search if the date is between 2010-06-04 and 2017-03-20.

- Using the ORDER BY keyword to sort the records by total number of landing, and using DESC keyword to sort the records in descending order.
- According to the results, the number of successes and failures between 2010-06-04 and 2017-03-20 was similar.

- Query

```
SELECT LANDING__OUTCOME,
       COUNT(LANDING__OUTCOME) AS total_number
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY total_number DESC
```

- Result

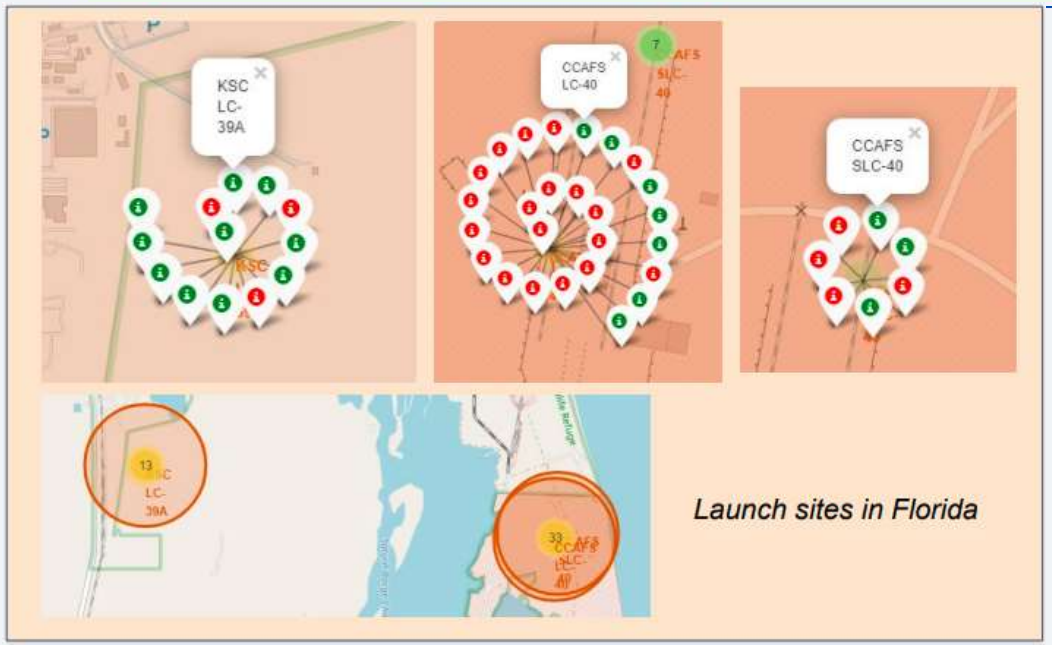
landing__outcome	total_
No attempt	
Failure (drone ship)	
Success (drone ship)	
Controlled (ocean)	
Success (ground pad)	
Failure (parachute)	
Uncontrolled (ocean)	
Precluded (drone ship)	

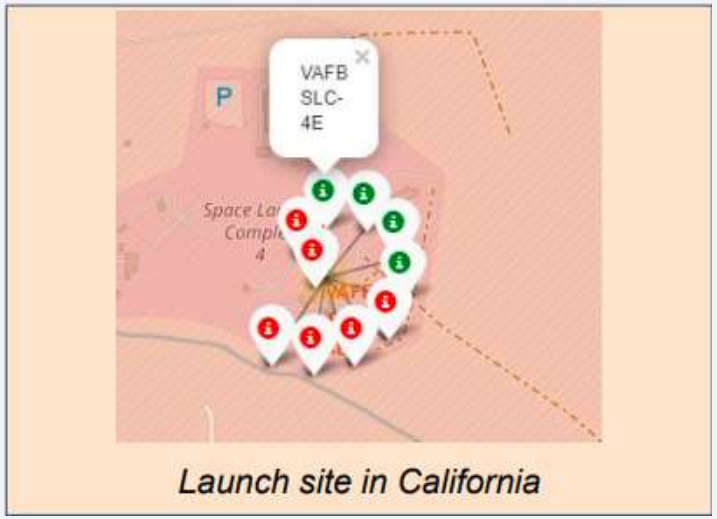

Section 3
Launch Sites
Proximities Analysis

All Launch Sites’ Locations

- The left map shows all SpaceX launch sites, and the right map also shows that all launch sites are in the United States.
- As can be seen on the map, all launch sites are near the coast.

Color-labeled Launch Outcomes



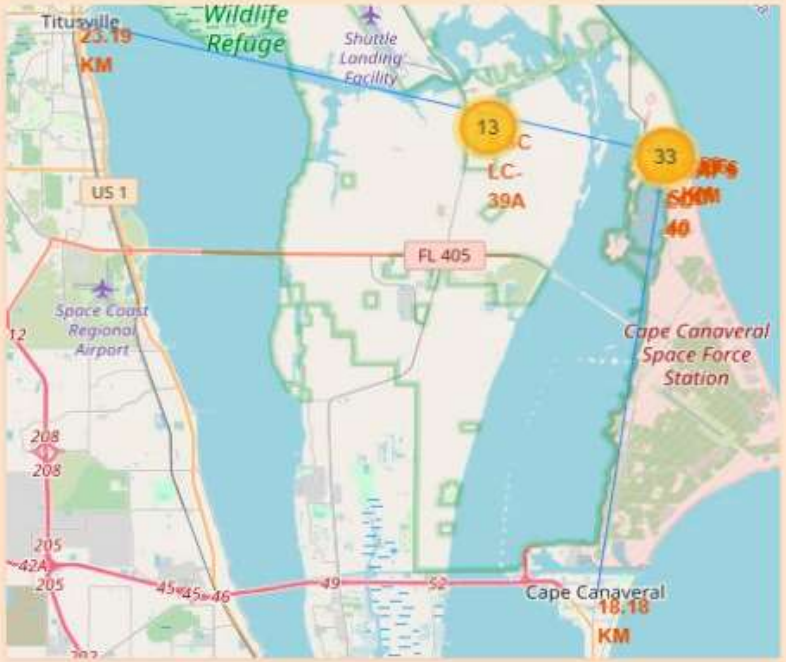


- By clicking on the marker clusters, successful landing (green) or failed lan (red) are displayed.

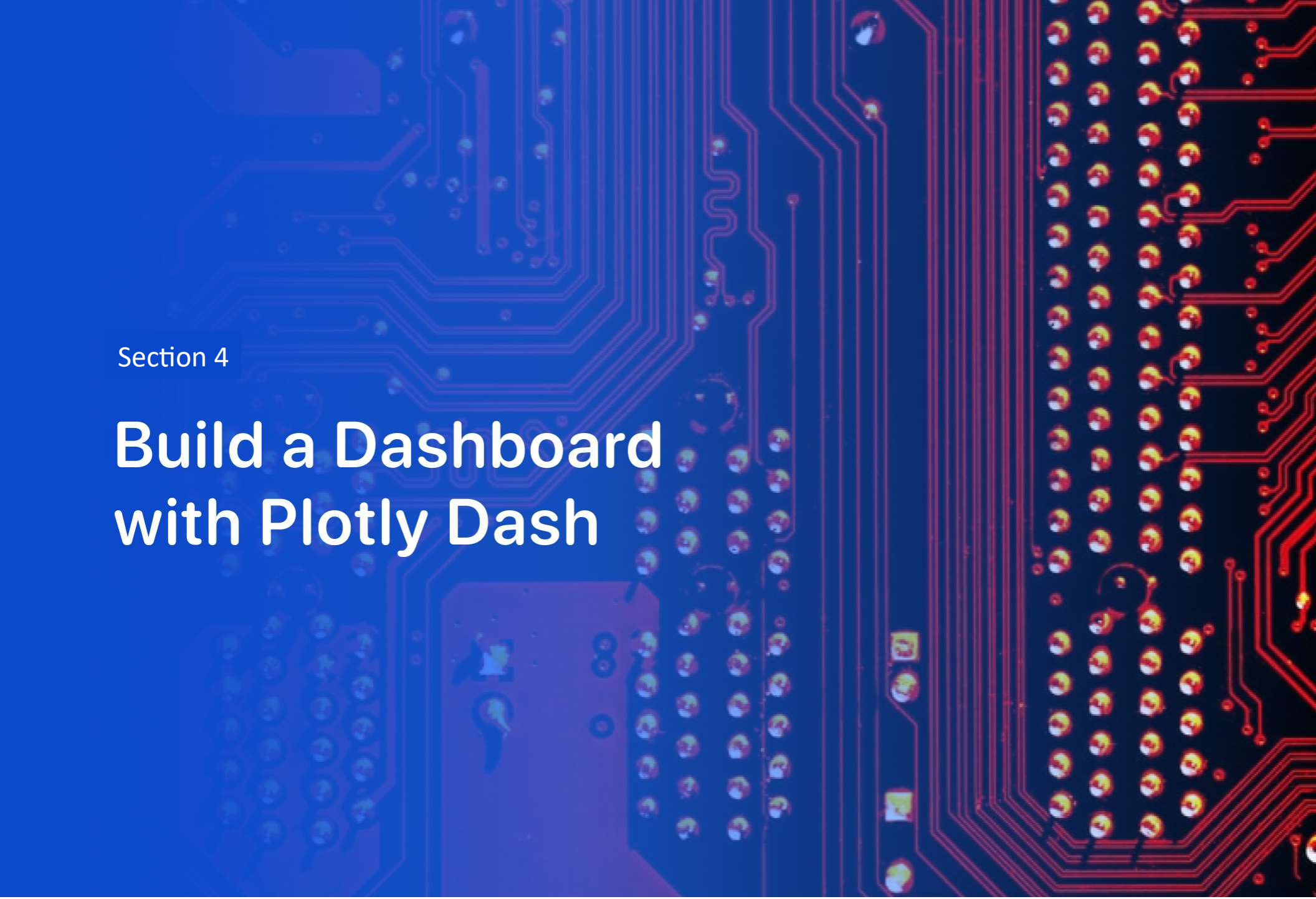
Proximities of Launch Sites

It can be found that the launch site is close to railways and highways for transportation of equipment or personnel, and is also close to coastline and relatively far from the cities so that launch failure does not pose a threat.

Are launch sites in close proximity to railways? ***Yes***
Are launch sites in close proximity to highways? ***Yes***
Are launch sites in close proximity to coastline? ***Yes***



Do launch sites keep certain distance away from cities? ***Yes***

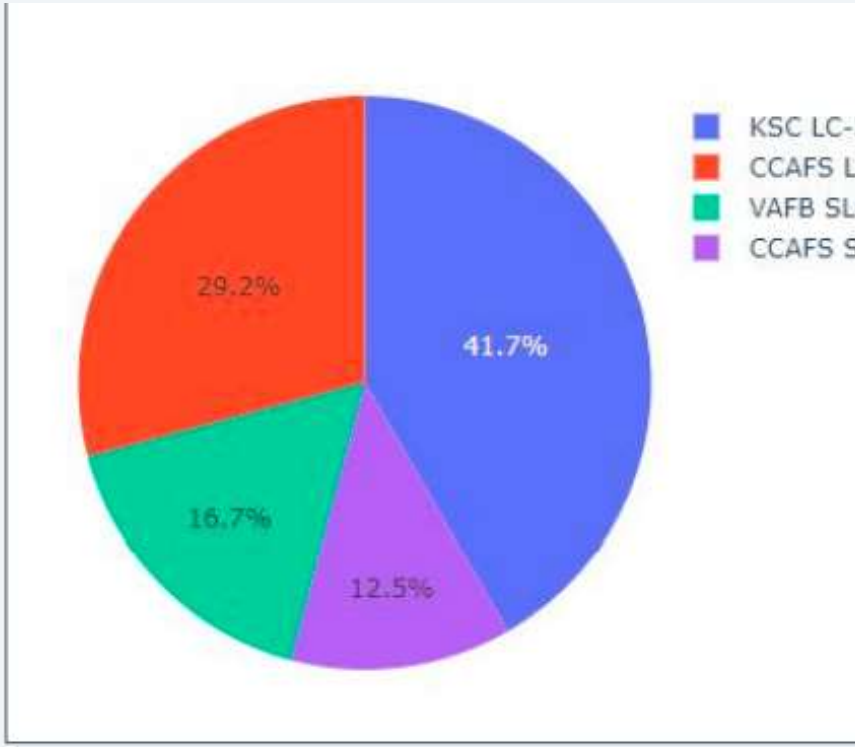


Section 4

Build a Dashboard with Plotly Dash

Total Success Launches By all sites

- KSLC-39A records the most launch success among all sites.
- The VAFB SLC-4E has the fewest launch success, possibly because
  - the data sample is small, or
  - because it is the only site located in California, so the launch difficulty on the west coast may be higher than on the east coast



Launch Site with Highest Launch Success Ra

- KSLC-39A has the highest success rate with 10 landing successes (76.9%) and 3

  landing failures (23.1%).

Payload vs. Launch Outcome Scatter Plot for all sites

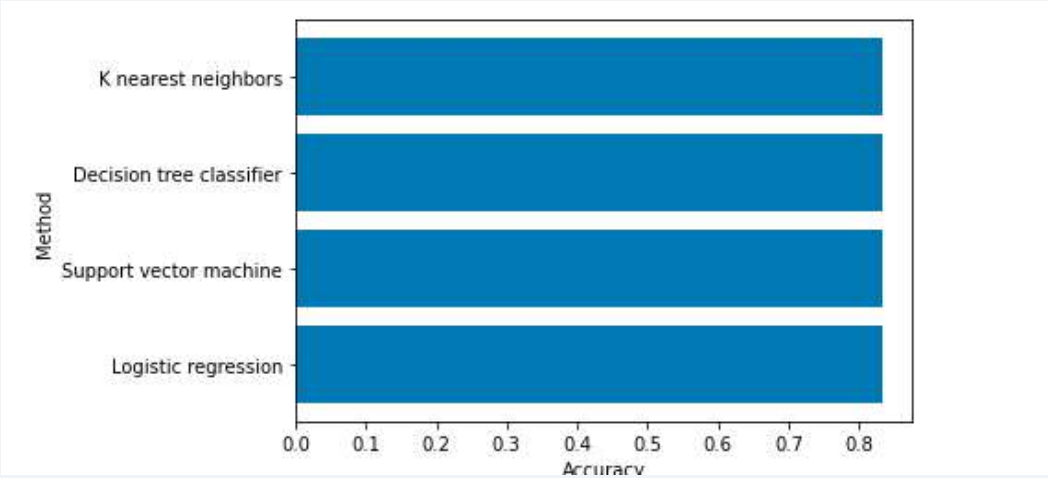
- These figures show that the launch success rate (class 1) for low weighted payloads(0-5000 kg) is hig
than that of heavy weighted payloads(5000-10000 kg).

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- After comparing the accuracy of the above methods, all return the same accuracy for the test data.

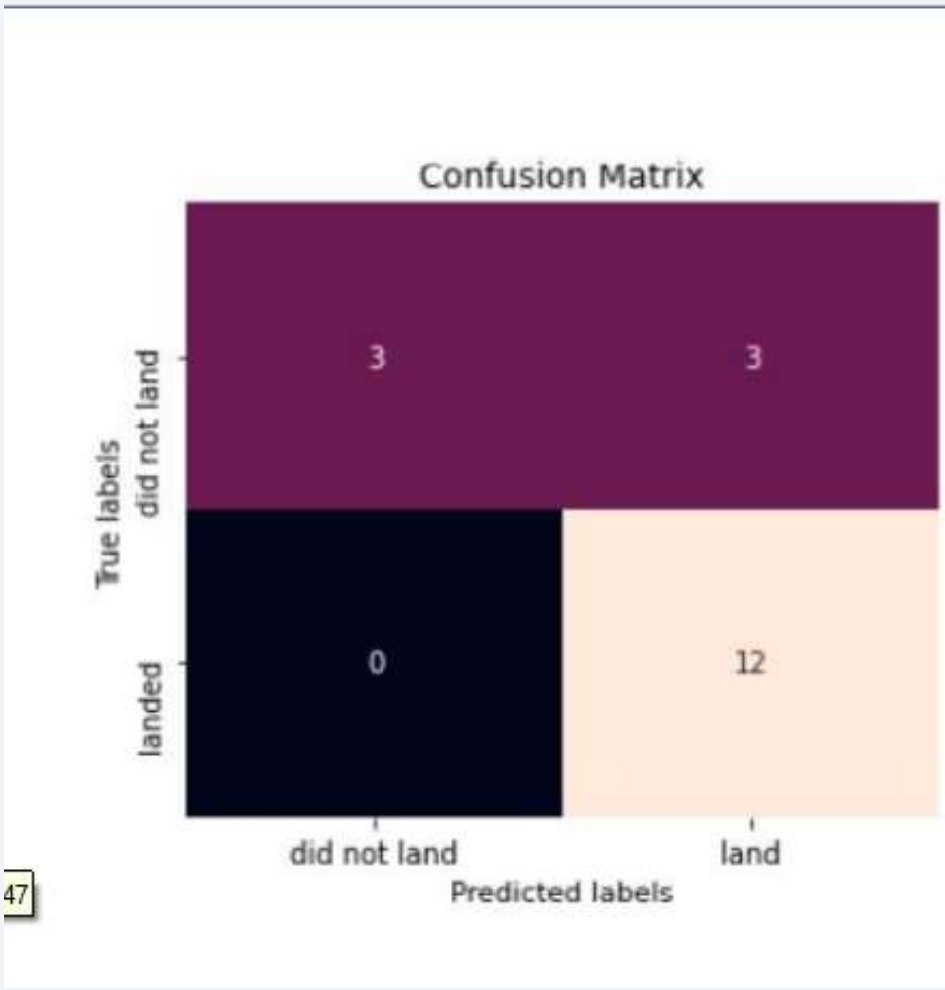


Confusion Matrix

- The confusion matrix is the same for all models because all models performed the same for the test set.

• The models predicted 12 successful landings when the true label was successful and 3 failed landings when the true label was failure. But there were also 3 predictions that said successful landings when the true label was failure (false positive).

• Overall, these models predict successful landings



Appendix

https://eu-de.dataplatform.cloud.ibm.com/analytics/notebooks/v2/83ef2ee2
e0-ac4a-3e96c536ff9a/view?access_token=f1331e75556f1144fb1195643
24b03aef81e9c8f9f338b8e8d2bbb3b

https://eu-de.dataplatform.cloud.ibm.com/analytics/notebooks/v2/64aa054
2e-b3ad-b88c3d3062e5/view?access_token=d83f6cc6e3905bfff04ea4e17
e23412f1bf76821b07745ad3849cfcb

https://eu-de.dataplatform.cloud.ibm.com/analytics/notebooks/v2/841f6eb2
14-b4c7-1c3ffc7a0d15/view?access_token=4228b75fcc381f7b45c9a68ce
5b4bccf586a937d0e7b0ac21072a5d79

Conclusions

- As the number of flights increased, the success rate increased, and recently has exceeded 80%.
- Orbital types SSO, HEO, GEO, and ES-L1 have the highest success rate (100%).
- The launch site is close to railways, highways, and coastline, but far from citi
- KSLC-39A has the highest number of launch successes and the highest succ rate among all sites.
- The launch success rate of low weighted payloads is higher than that of heav weighted payloads.
- In this dataset, all models have the same accuracy (83.33%), but it seems th more data is needed to determine the optimal model due to the small data siz

Thank you!