IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Brayan Christian
11/25/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Relevant Data is collected from multiple trusted sources like SpaceX REST API and Wikipedia.
  - Exploratory analysis of different factors is done using SQL and Python to answer questions like, how factors like launch sites and payload affect the success of a mission.
  - Visualization of various statistical tests to understand the significance of some features, to understand the reason for success rate of particular rockets.
  - Machine learning models are developed to predict the success of a launch from factors like launch site, etc.
- Summary of all results
  - There is an innate relationship between successful launch and launchsite, payload mass, etc.
  - Machine learning algorithms that predicts the success of a launch based on said factors with 83% accuracy.

Introduction

- SpaceX is turning heads with its revolutionary reusable rockets, revolutionizing the industry. This feat, of using reusable rockets, has decreased the cost of space travel by about a 100 million.
- However, while this revolutionary idea is fascinating, there exists a dark shadow of doubt regarding the ability of SpaceX's rockets to land successfully.
- Through this analysis, I hope to find:
  - The factors that contribute to a successful mission (take-off and landing).
  - Analysis of how significant a factor is to a mission's success rate.
  - Consistency rate with which SpaceX is able to land their spacecraft.
  - Ability to predict if a given spacecraft, under certain circumstances and factors will be able to do a successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data is collected from trusted sources like SpaceX REST API and Wikipedia.
- Perform data wrangling
  - Clean the data, replace missing values, make the data normal distributed and perform encoding.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Multiple classification algorithms are used with different parameters to find the best hyperparameters.

Data Collection

- Data is collected from trusted sources which include:
  - SpaceX REST API: https://api.spacexdata.com/v4

    This API provides the data of records from all launches including rockets used, launch dates, payload masses, outcome, and many more (over 47 features).

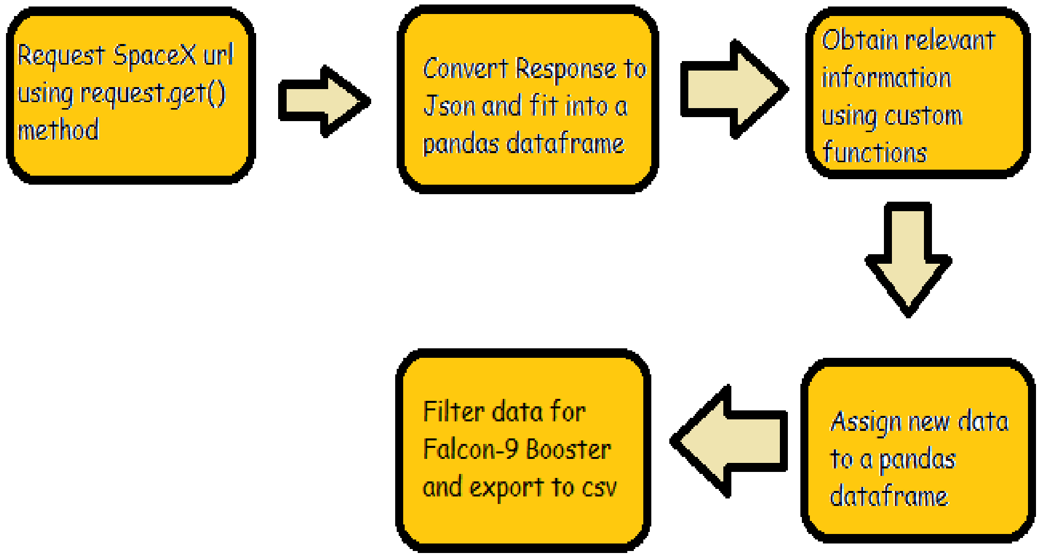
    Since this is the official API of SpaceX, the data can be trusted. However, there might be some nuances that might be missing from this data due to various factors. To mitigate this,
  - Wikipedia: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

    Data regarding Falcon 9 rockets is extracted from this webpage using Beautiful Soup.

    This data contains all the history and factors surrounding the Falcon 9 rocket launches.

Data Collection – SpaceX API

- Data is collected from get request method. From this only the relevant data is extracted using specialized functions.
- Data is cleaned and fitted to a pandas dataframe.
- Data is filtered to contain only Falcon-9 booster.
- Github: https://github.com/brayan-edison/SpaceX-Capstone/blob/main/jupyter_labs_spacex_data_collection_api.ipynb



Request SpaceX url using request.get() method

Convert Response to Json and fit into a pandas dataframe

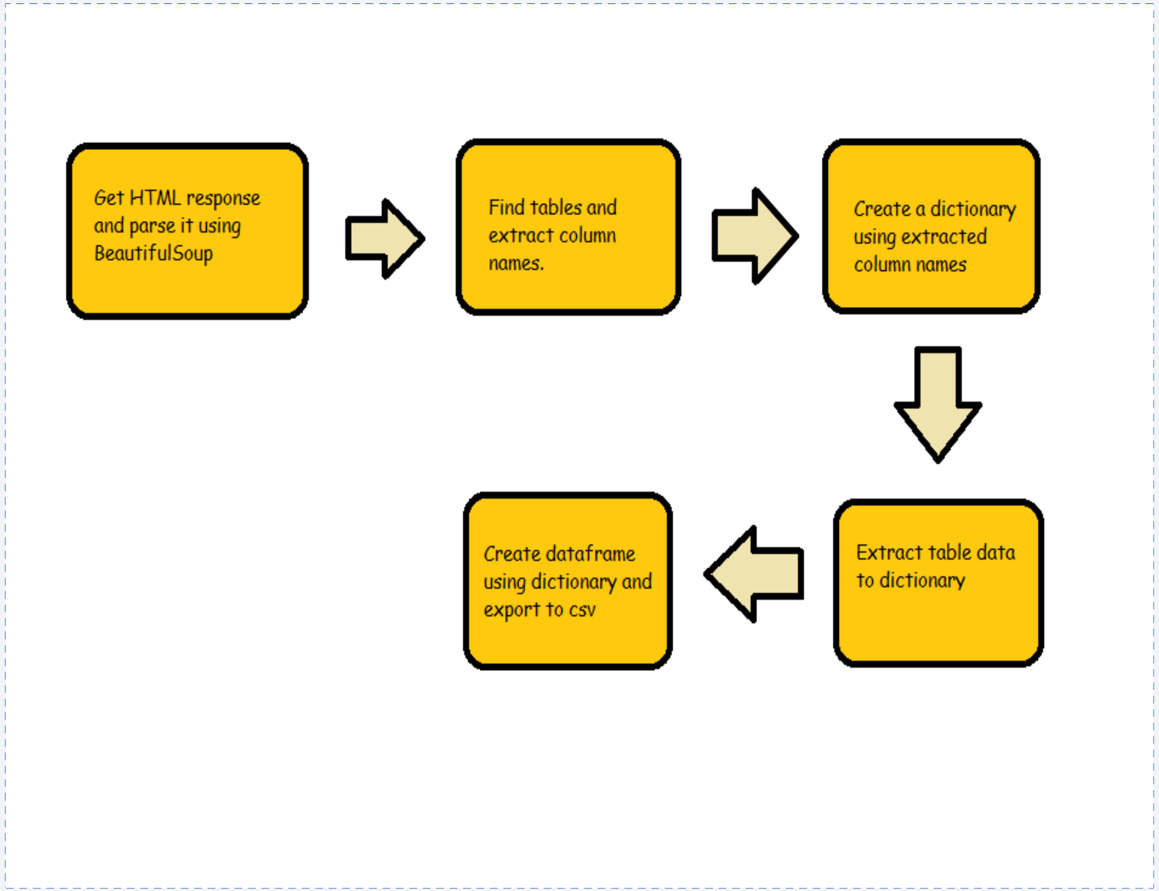
Obtain relevant information using custom functions

Assign new data to a pandas dataframe

Filter data for Falcon-9 Booster and export to csv

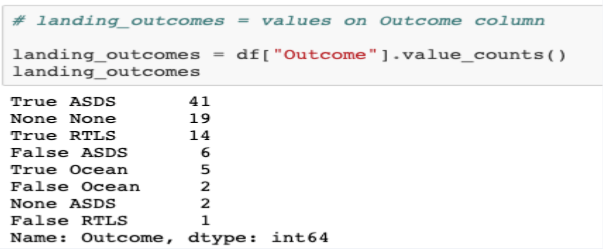
Data Collection - Scraping

- Create HTML response using BeautifulSoup.
- Find tables and extract column names to a dictionary.
- Extract data from tables to the dictionary and create a pandas dataframe from that dictionary.
- Github: https://github.com/brayan-edison/SpaceX-Capstone/blob/main/jupyter_labs_webscraping.ipynb


Get HTML response and parse it using BeautifulSoup
Find tables and extract column names.
Create a dictionary using extracted column names
Extract table data to dictionary
Create dataframe using dictionary and export to csv

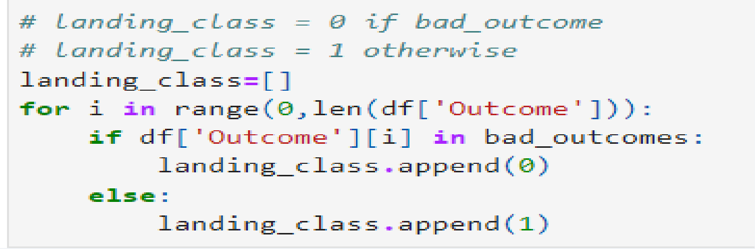
Data Wrangling

- Feature engineering was used to transform outcome feature from multiple categorical values to binary success or failure.



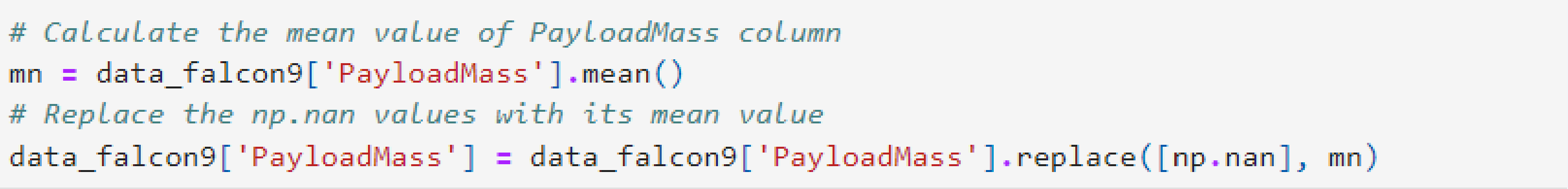
```
# landing_outcomes = values on Outcome column

landing_outcomes = df["Outcome"].value_counts()
landing_outcomes
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```



```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class=[]
for i in range(0,len(df['Outcome'])):
    if df['Outcome'][i] in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
```

- Data cleaning to remove all null values by replacing them with the mean.



```
# Calculate the mean value of PayloadMass column
mn = data_falcon9['PayloadMass'].mean()
# Replace the np.nan values with its mean value
data_falcon9['PayloadMass'] = data_falcon9['PayloadMass'].replace([np.nan], mn)
```

- Github: https://github.com/brayan-edison/SpaceX-Capstone/blob/main/labs_jupyter_spacex_Data_wrangling.ipynb

EDA with Data Visualization

- Bar Graphs:
    - To understand the success rate based on orbit radius.
- Scatter Plot:
    - To understand the relationship between multiple features and the outcome.
    - To understand the correlation and linearity of features against one another.
- Line Graph:
    - To visualize the landing success rate with respect to time. To understand if the efficiency has increased.

Github: https://github.com/brayan-edison/SpaceX-Capstone/blob/main/jupyter_labs_eda_dataviz.ipynb

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA‘
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass
- List the failed landing_outcomesin drone ship, their booster versions, and launch site names for in year 2015
- 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

Github: https://github.com/brayan-edison/SpaceX-Capstone/blob/main/jupyter_labs_eda_sql_coursera.ipynb

Build an Interactive Map with Folium

- Launch site coordinates were used to mark the location of every launch with a circle marker.
- Used green marker for every successful launch and red for every failed launch/mission.
- Measured distance between launch sites and nearest coastline, city, highway and railways using a line.
- Questions answered using Folium analysis:
  - Is there a significance between proximity of launch site with neighboring features?
  - How likely does a launch site and its proximity region affect the outcome?
- Github: https://github.com/brayan-edison/SpaceX-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

Anatomy of Dashboard:

- Pie Chart of outcome based on launch site:
  - Pie chart was used to get a clear idea about the difference in success rate between different launch sites.
- Scatter Plot of Payload Mass vs. Landing Success Rate:
  - This sheds light on the relation between payload mass and success rate.
  - A scaler to select the range of payload mass to understand the relationship better.

Github: https://github.com/brayan-edison/SpaceX-Capstone/blob/main/Dash_Spacex.ipynb

Predictive Analysis (Classification)

Build:

- Perform Data Wrangling and split it into train, validation and test datasets.
- Create an ensemble of classification algorithms (KNN, SVM, Logistic Regression, Decision Tree).
- Train the models with multiple hyperparameters using gridsearch to find the best accuracy on training and validation dataset.

Evaluation and Selection:

- Test the model on the Test dataset and plot confusion matrix to find Type 1 and Type 2 errors.
- Select the algorithm that has the lowest Type 1 error, however both are bad and need to have a balance.

Github: https://github.com/brayan-edison/SpaceX-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.ipynb

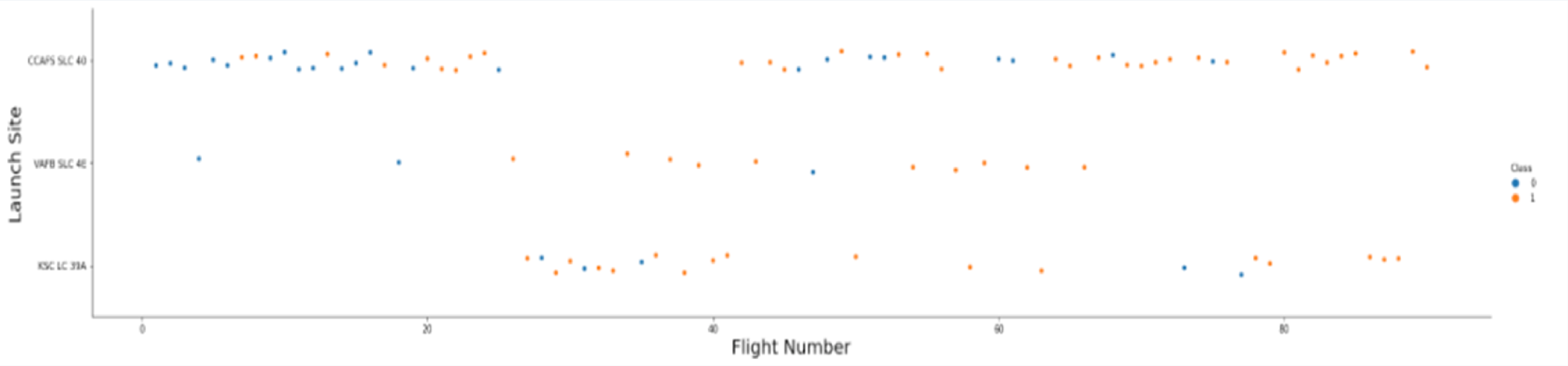
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

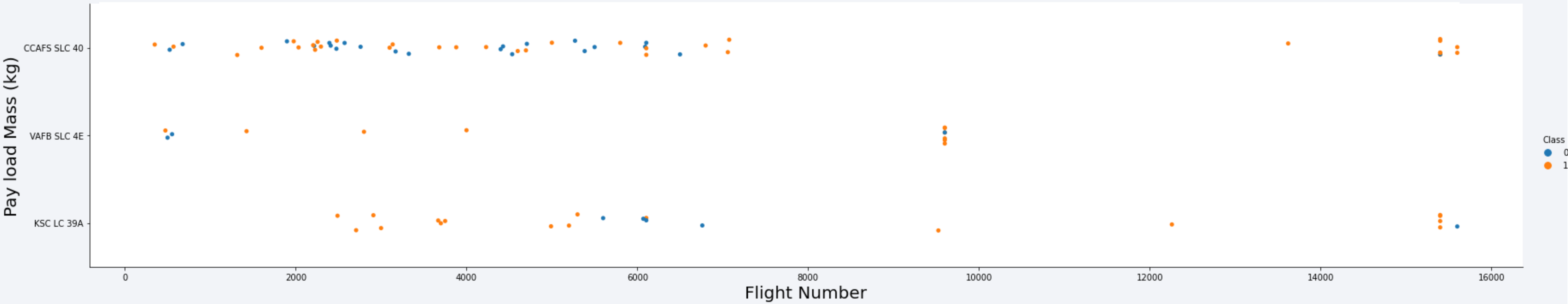
Insights drawn from EDA

Flight Number vs. Launch Site



- It can be observed that the success rate increases with flight number.
- It can also be observed that the success rate of launch site KSC LC 39A has the highest success rate.

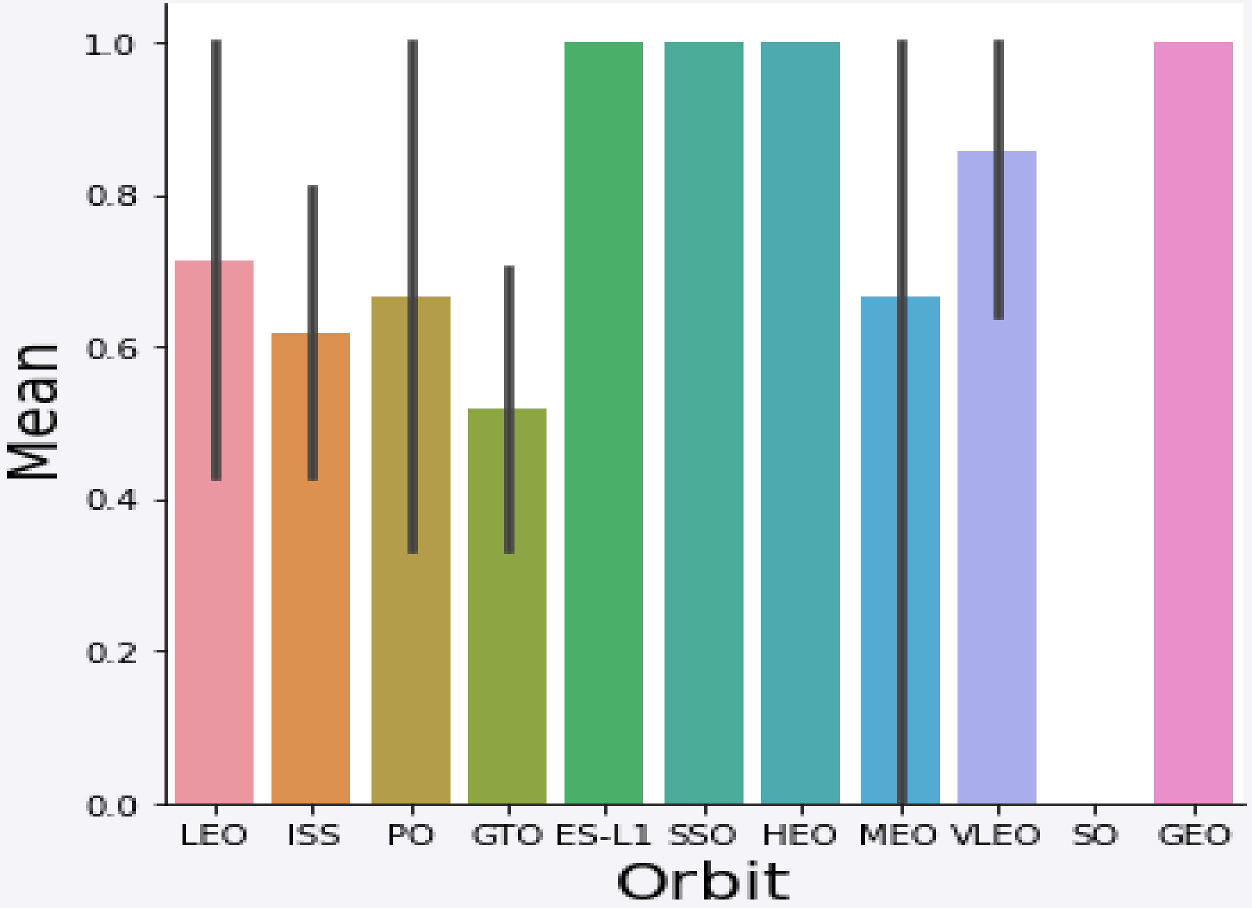
Payload vs. Launch Site



- It can be clearly observed that the success rate increases with increase in Payload mass.
- Another observation is that, for mid-range payload, VAFB SLC 4E site is used, which other two sites are used for heavy-range payload.

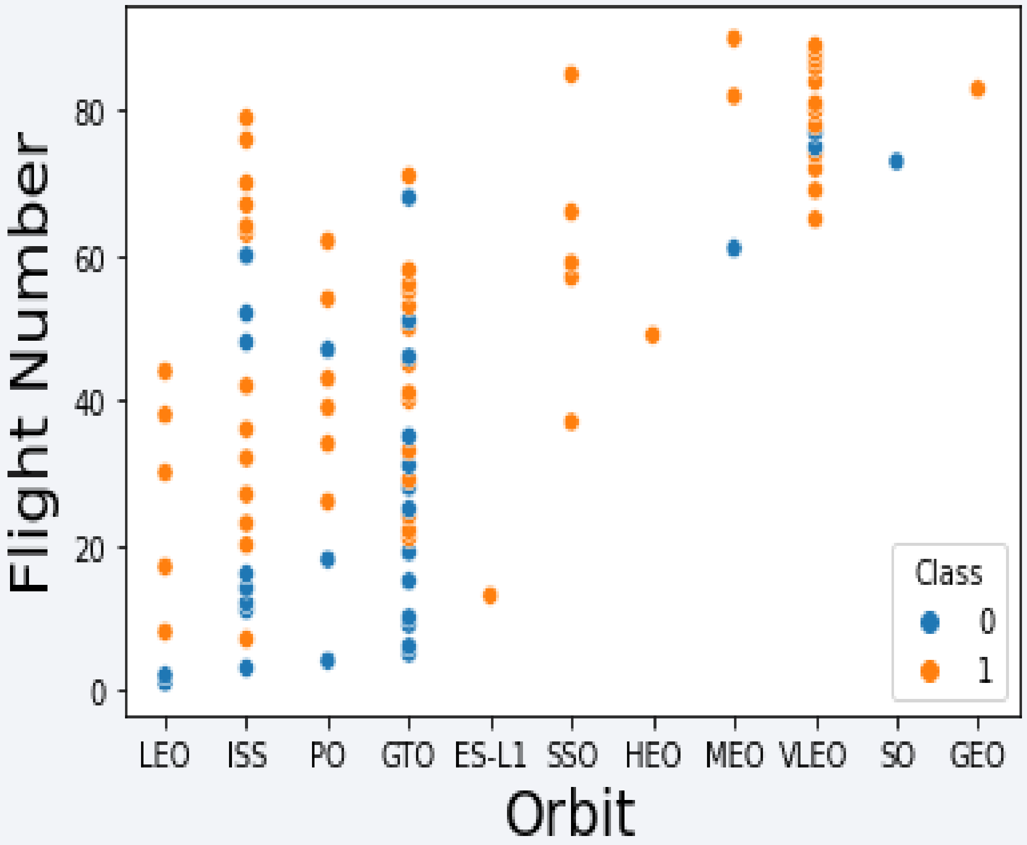
Success Rate vs. Orbit Type

- We can deduce from the barplot that ES-L1, SSO, HEO, GEO satellite ranges have perfect success rate.
- The poorest performing launch range is GTO with 50% chance of success.
- It can be deduced that low and high altitude launches are more successful compared to mid-altitude launches.



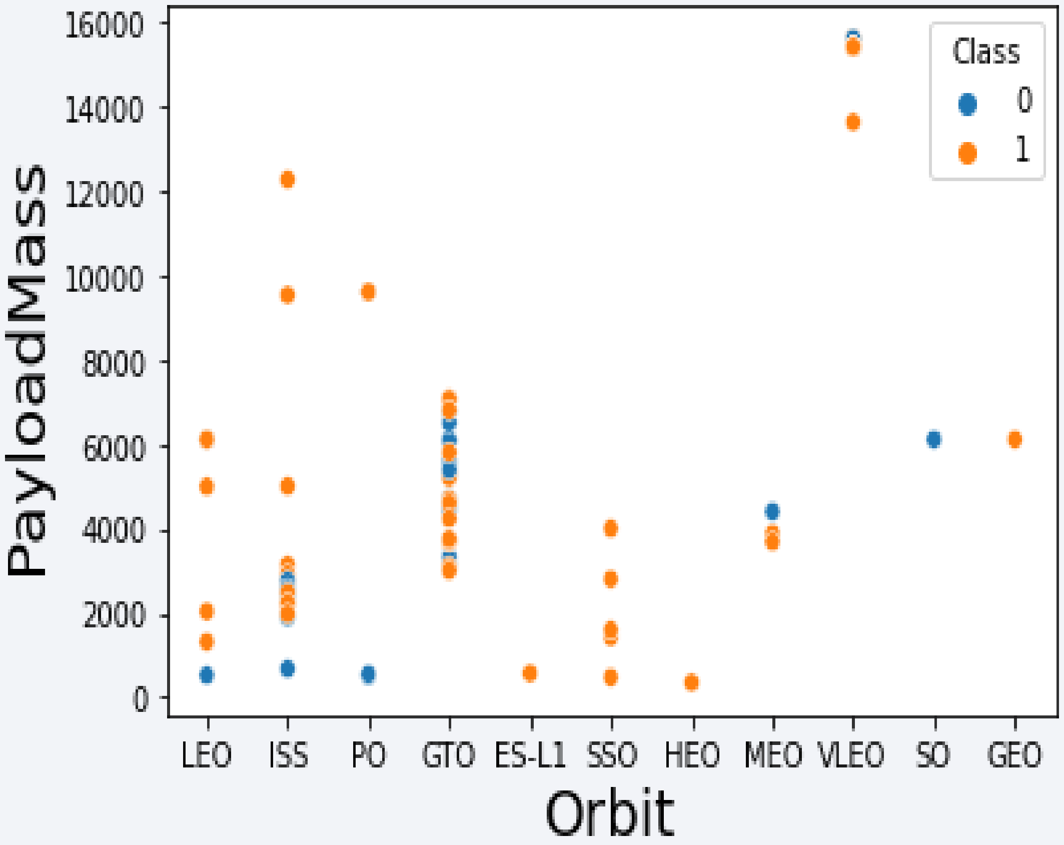
Flight Number vs. Orbit Type

- There is a clear relationship that with every increase in flightnumber, the success rate has improved for all orbit ranges.
- Apart from some initial failures, LEO orbit success rate has consistently improved. While ES-L1, SSO, HEO and GEO have perfect success rate but that may be due to less frequency of launches.
- GTO is where SpaceX really seems to struggle as the success rate is quite low. There has been no improvement it seems in making the success viable in this orbit range.



Payload vs. Orbit Type

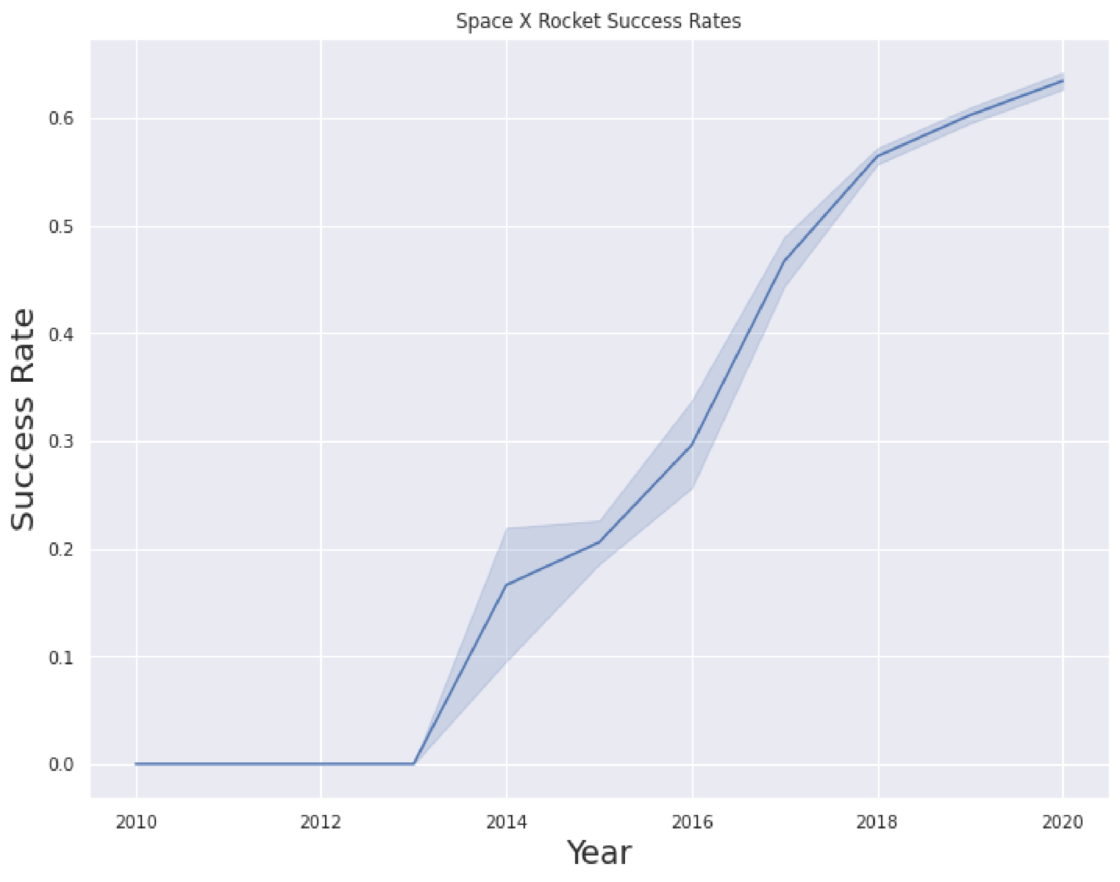
- It seems that lower mass payloads are mostly for LEO, ISS, PO, ES-L1, SSO and HEO.
- Most midrange mass payloads are for GTO and indeed most frequent orbit aim is GTO.
- There are not enough flights, apart from GTO and ISS to properly understand the relationship between payload mass and orbit.



Launch Success Yearly Trend

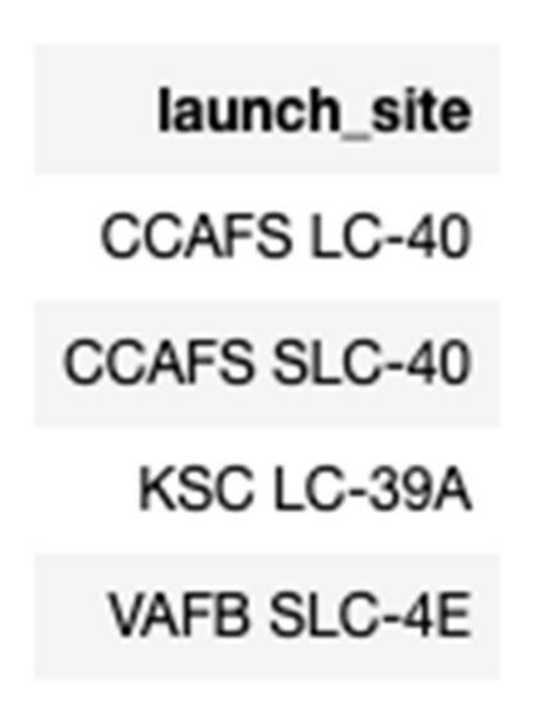
- There is a general increase in success rate since 2013.

- It also seems that the rate of increase in success is saturating, i.e. the rate of increase in success rate is increasing but with a flatter slope from 2018.



All Launch Site Names

- SQL magic is used to query db2 database.
- Query: %sql SELECT UNIQUE launch_site from SpaceX
- By fetching unique values of launch sites, we can narrow our Approach to group launches.

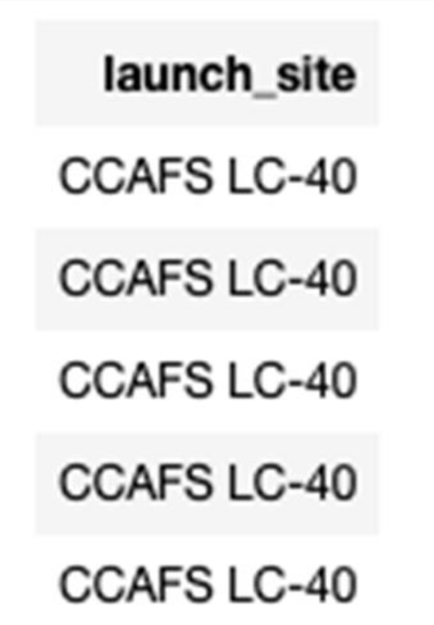


launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Query: %%sql Select * from SpaceX
  where Launch_Site like 'CCA%‘ limit 5

- Explanation:
  - The above query returns every launch site that begins with ‘CCA’
  - The where clause limits the result with launch sites that begin with ‘CCA’



launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

Total Payload Mass

- Query:

```
%%sql Select customer, sum(payload_mass_kg) as "Total Payload Mass" from
(Select customer, payload_mass_kg from SpaceX
where customer LIKE 'NASA (CRS)')
GROUP BY CUSTOMER
```

Customer	Total Payload Mass
NASA (CRS)	45596

- Explanation:
  - This query uses a subquery.
  - The subquery retrieves customer and payload mass where customer is NASA (CRS)
  - The main query then calculates the sum of payload mass and gives the result.

Average Payload Mass by F9 v1.1

- Query:

```
%%sql Select AVG(payload_mass_kg) as "Average
Payload Mass" from SpaceX where booster_version LIKE ‘F9 v1.1%’
```

- Explanation:
  - The query selects the average of payload mass where booster version is F9 v1.1

Average Payload Mass
2928

First Successful Ground Landing Date

- Query:

```
%%sql Select MIN(DATE) as "First Success" from
(select DATE from SpaceX
where LANDING_OUTCOME LIKE 'Success%')
```

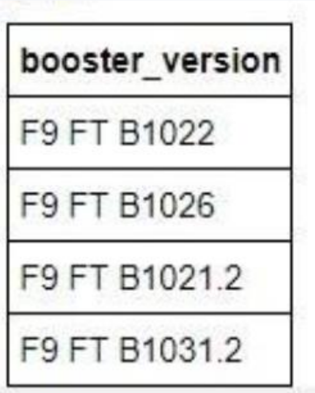
First Success
2015-12-22

- Explanation:
  - The query selects the minimum of data, i.e. earliest of dates from the subquery.
  - The subquery returns all dates where landing outcome was successful.

Successful Drone Ship Landing with Payload between 4000 and 6000

- Query:

```
%%sql SELECT BOOSTER_VERSION
from SpaceX
where (payload_mass_kg Between 4000 and 6000)
and (Landing_Outcome = 'Success (drone ship)')
```



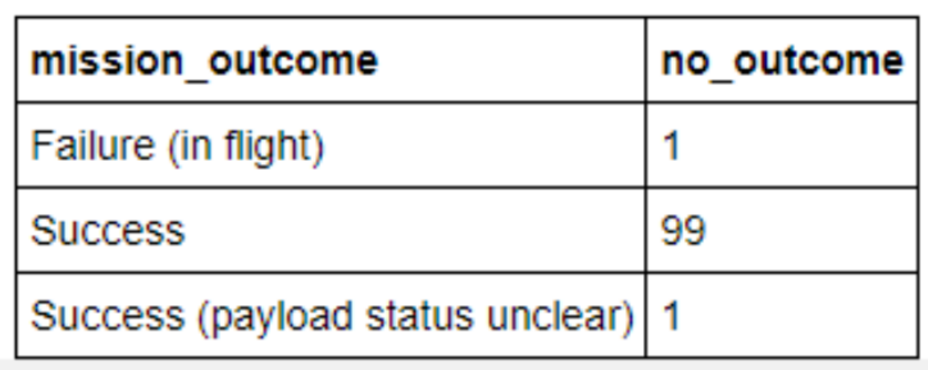
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- Explanation:
  - SQL query to find booster version of all instances where payload mass was between 4000 and 6000 and the mission was a success.

Total Number of Successful and Failure Mission Outcomes

- Query:

```
%%sql SELECT mission_outcome, count(mission_outcome) as “no_outcome"
from
SpaceX
Group by mission_outcome
```



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

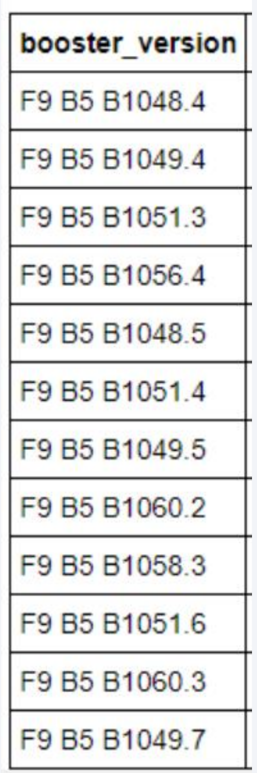
- Explanation:
  - The SQL query finds all mission outcome and the sum of all mission outcome with the alias no_outcome.

Boosters Carried Maximum Payload

- Query:

```
%%sql SELECT Unique booster_version from SpaceX
where payload_mass_kg = (Select max(payload_mass_kg) from SpaceX)
```

- Explanation:
  - The query selects the booster version where payload mass is maximum.

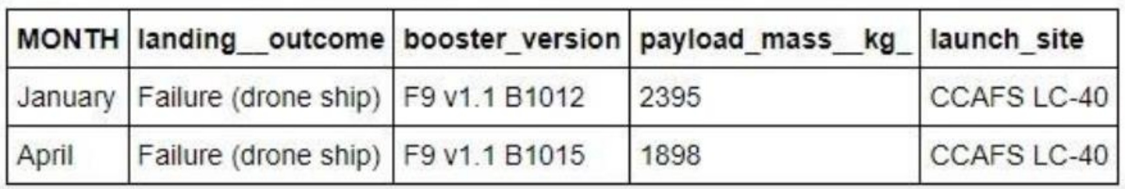


booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

2015 Launch Records

- Query:

%%sql Select MONTH, landing_outcome, booster_version, launch_site from SpaceX
where landing_outcome = 'Failure (drone ship)' and Year(DATE) = 2015



MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

- Explanation:
  - The query selects the month, landing outcome, booster version and launch site where landing outcome is a failure and year is 2015

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

- Query:

```
%%sql select landing_outcome, count(landing_outcome)
as "Total"
from Spacex
where DATE between '2010-06-04' and '2017-03-20‘
group by landing_outcome
order by "Total" desc
```

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

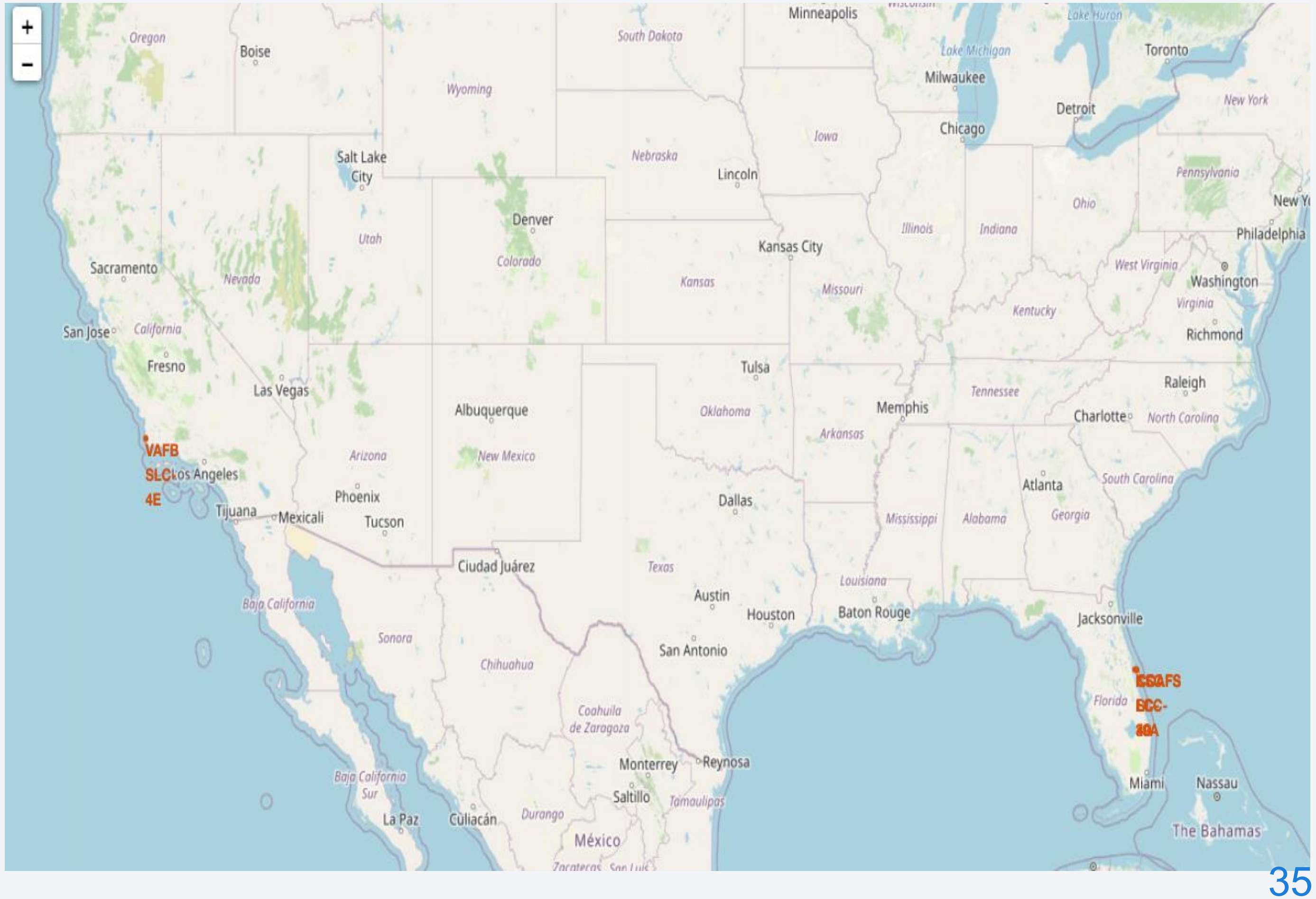
- Explanation:
  - The query selects landing outcome and their count as total between the date range of ‘2010-06-04’ and ‘2017-03-20’ which is grouped by landing outcome and is displayed in descending order.

Section 3

Launch Sites Proximities Analysis

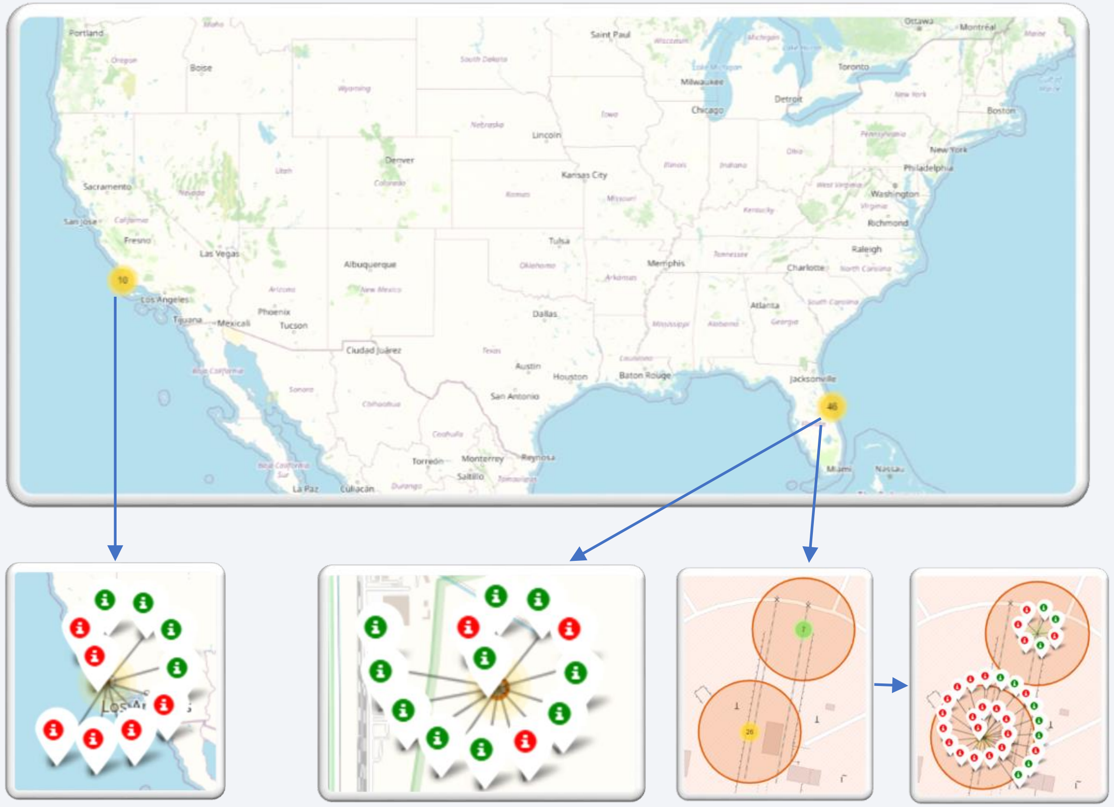
Analysis of Launch Sites for SpaceX missions:

- Inference:
  - All of SpaceX launch sites are located in USA.
  - It utilizes 4 sites, from which 3 are located on the East Coast in Florida. While only one is located in the West Coast.
  - All of the launch sites are in close proximity to coastline.



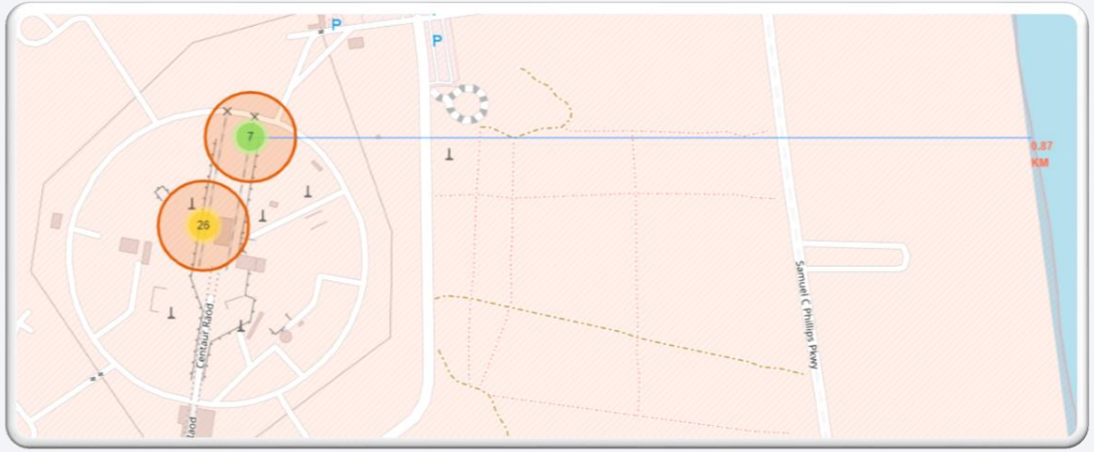
Analysis of success and failure for SpaceX missions:

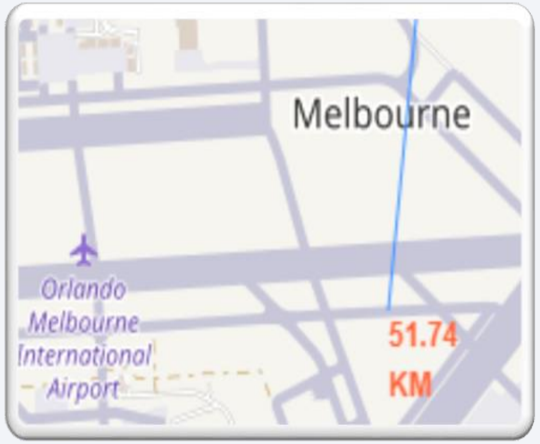
- As can be seen from the map, most of the launches have been done from the east coast.
- The success rate of west coast is 40% while it is higher in some launch sites in the east coast.

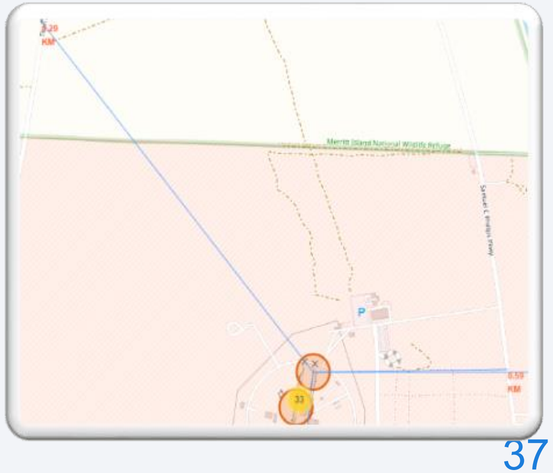


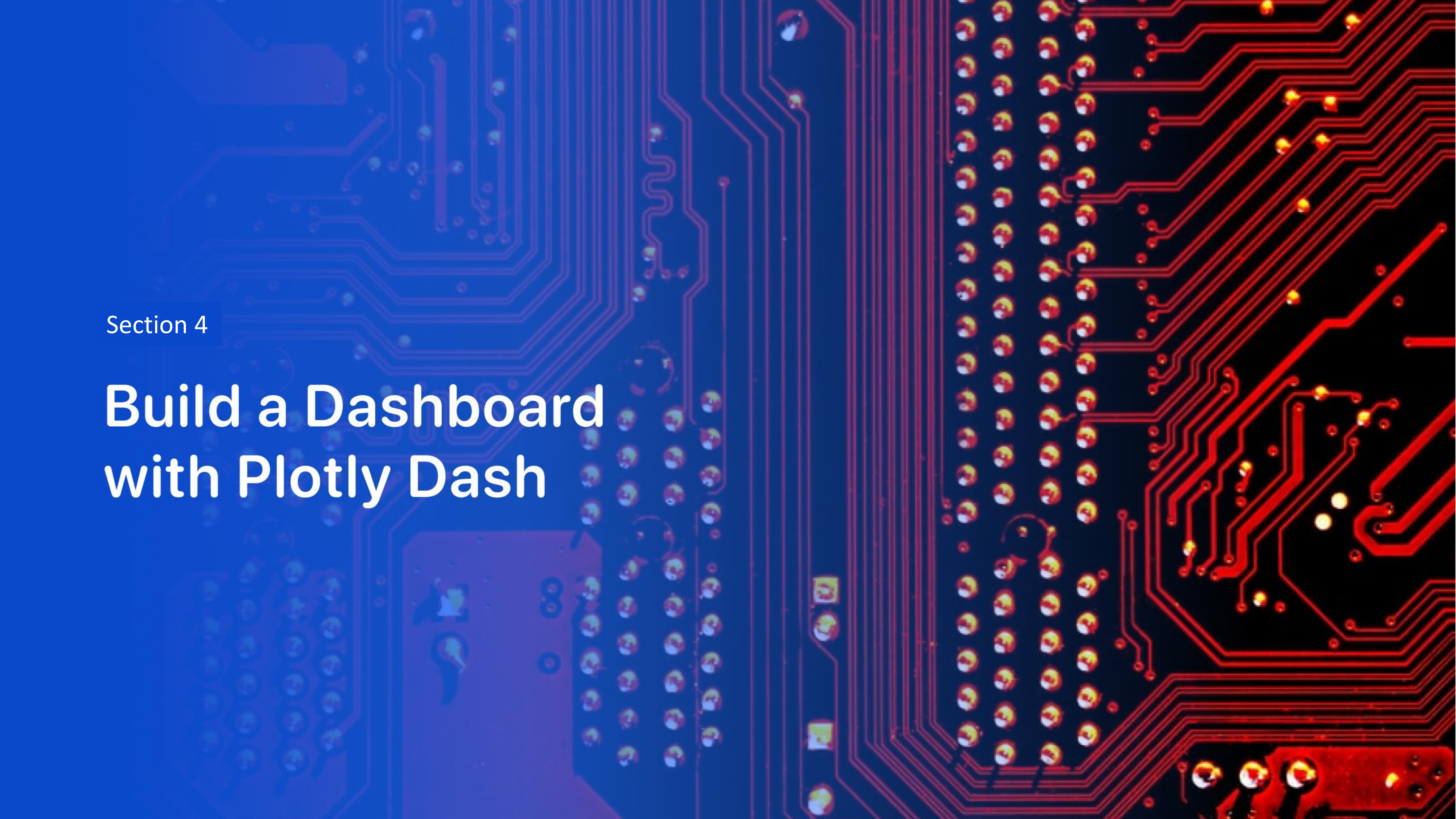
Proximity Analysis of Launch Sites

- It can be noticed that launch sites are in close proximity with railways and highways.
- The nearest highway is at a distance of 0.59 km while the nearest railway is at a distance of 1.29 kms.
- It is most certainly intentional as it would be easier to transport components from different parts.







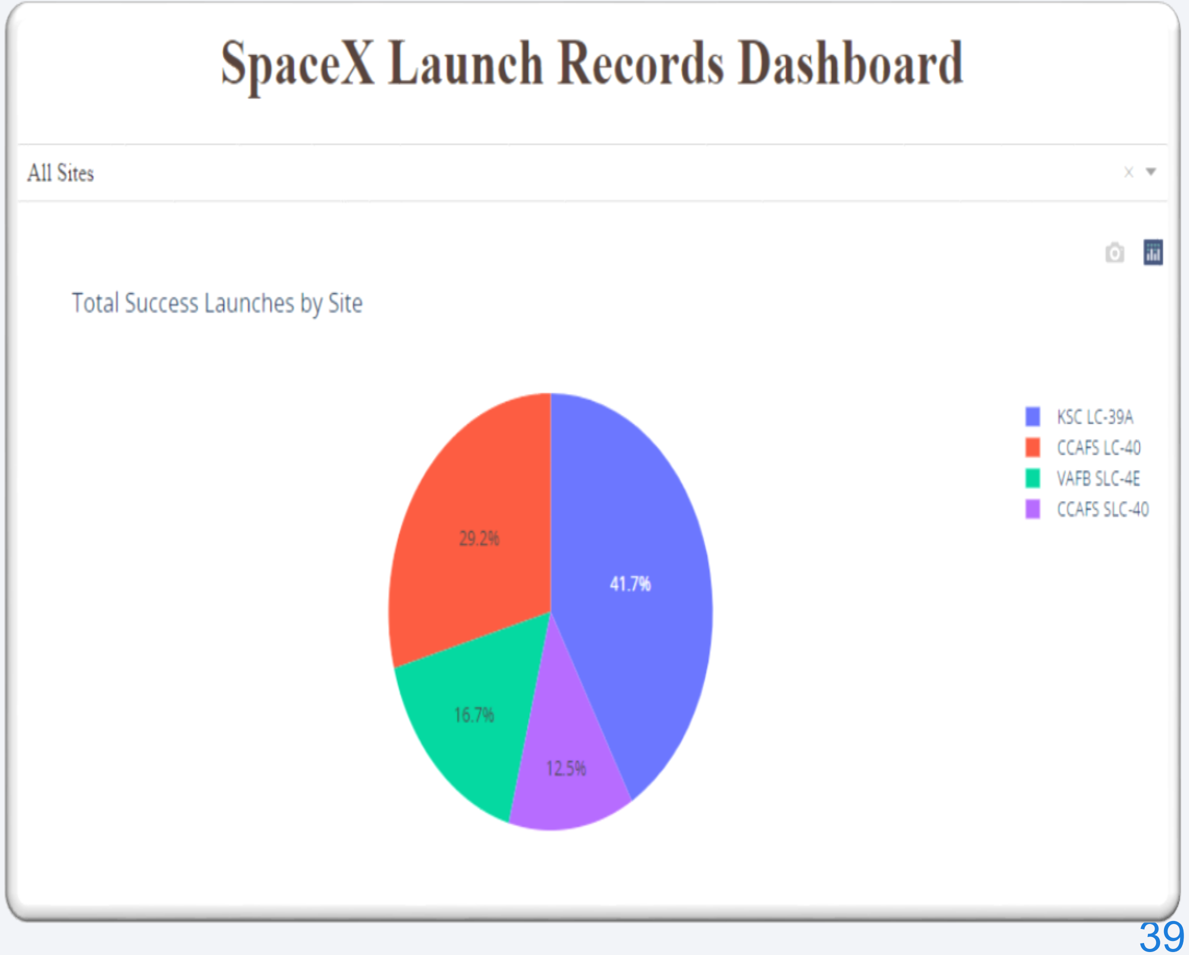


Section 4

Build a Dashboard with Plotly Dash

SpaceX Success Records

- The dashboard hosts a pie chart of success rates for each launch site.
- It is clear that KSC LC-39A has been most successful landing rate, while CCAFS SLC-40 has been the poorest performer.

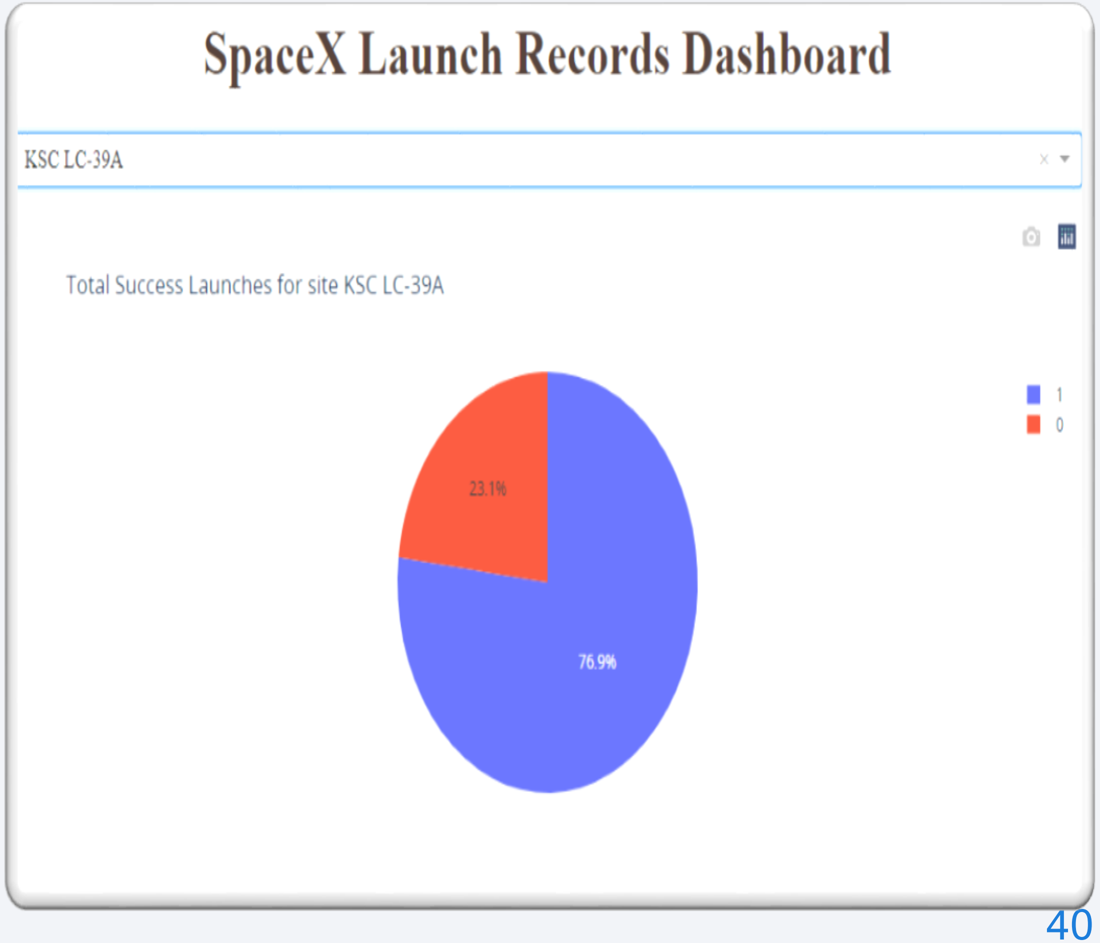


SpaceX Launch Records Dashboard

All Sites

In-Depth look at site KSC LC-39A

- Taking an in-depth look at the most successful launch site for SpaceX, it is found that it has 76.9% success rate of landing.

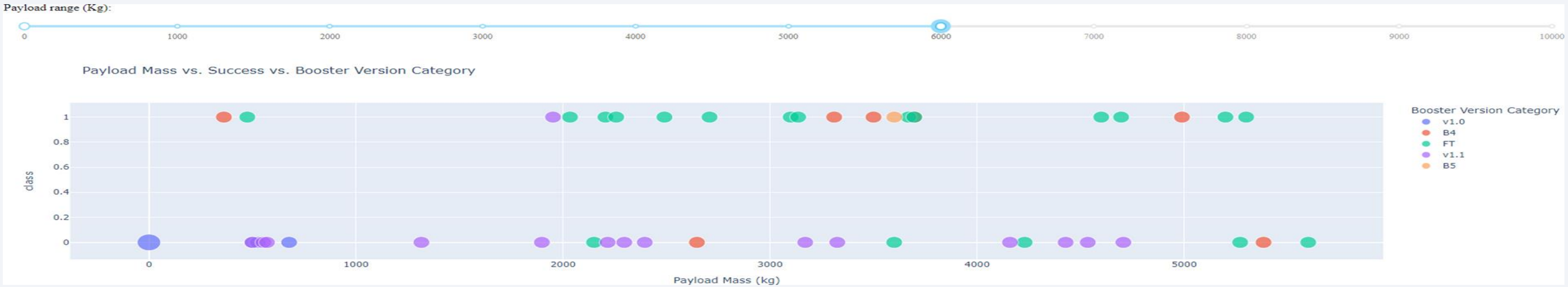


SpaceX Launch Records Dashboard

KSC LC-39A

Total Success Launches for site KSC LC-39A

Relationship between Payload mass and success



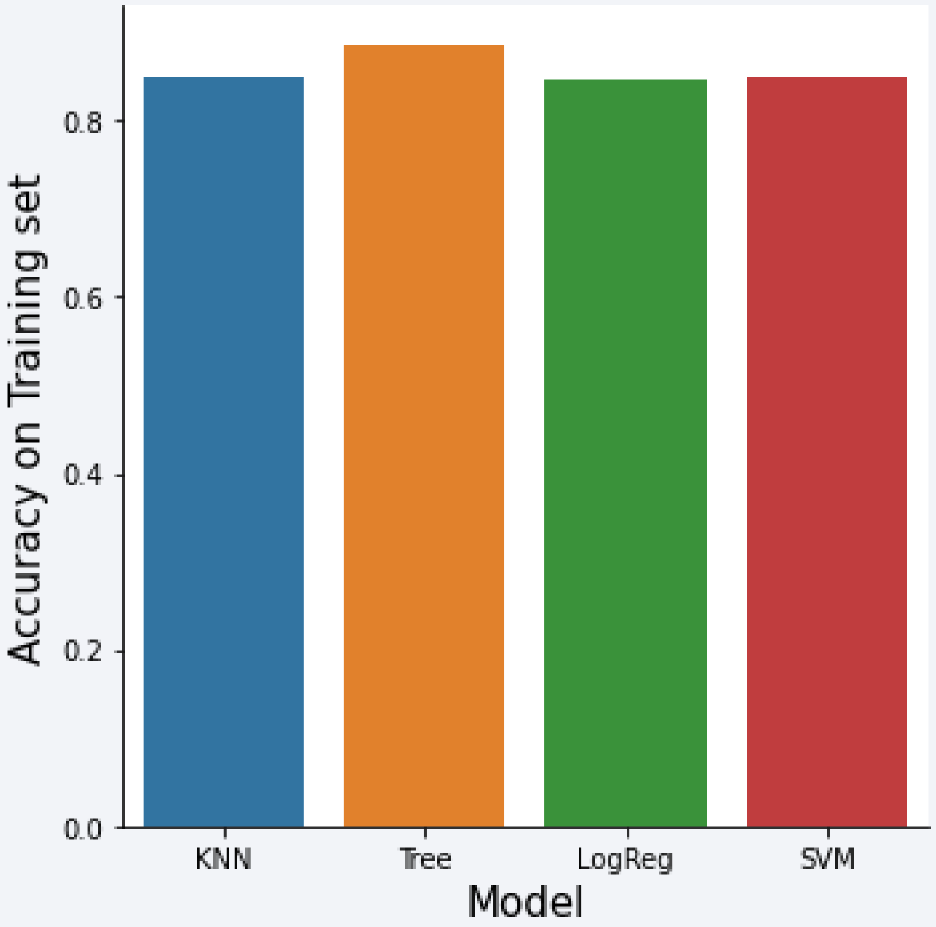
- Upon looking at the relationship between payload mass and success rate, it can be found that the booster FT has the highest success rate while v1.1 performs the worst.
- Most of payload are in mid-range (2000-4000kg).

Section 5

Predictive Analysis (Classification)

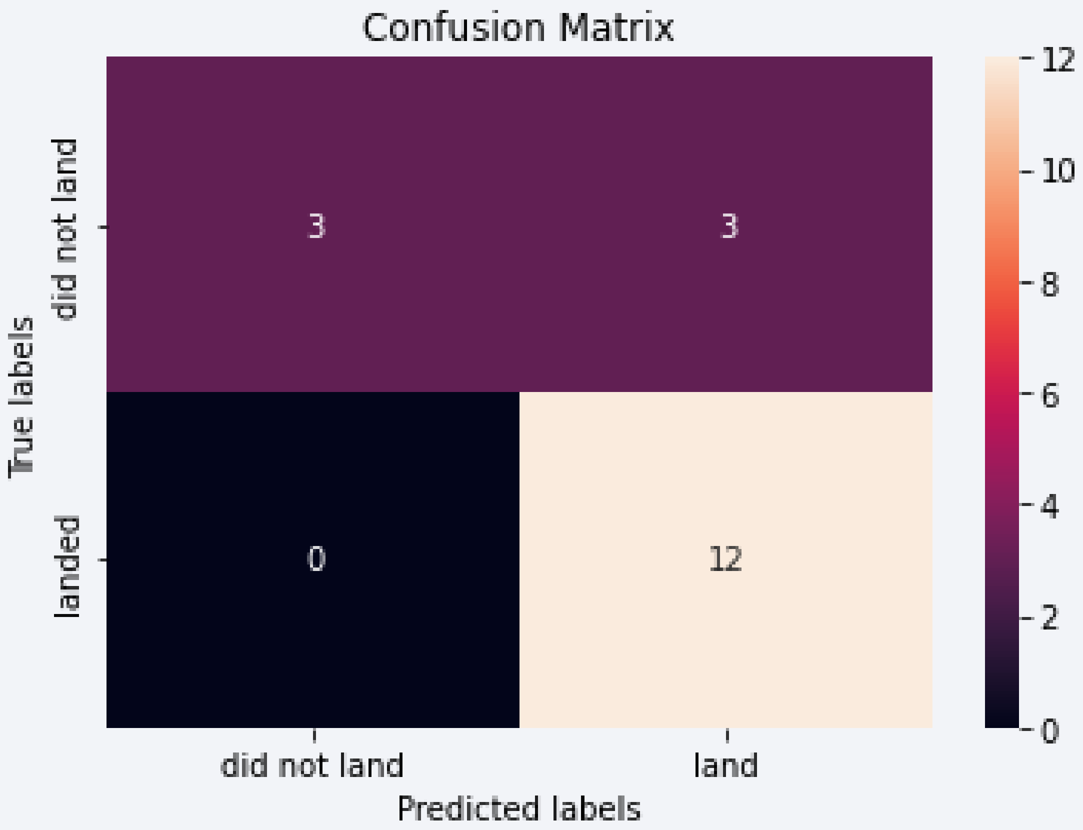
Classification Accuracy

- Decision tree has the highest training and testing accuracy from the ensemble of models KNN, Logistic Regression, SVM and Decision tree.



Confusion Matrix

- The decision tree model predicts perfectly if a mission will have a successful landing.
- The model throws false positives, a type 1 error.



Conclusions

- The success rate has increased significantly from the earlier years but it seems it is reaching saturation.
- Launch site KSC LC-39A has the highest success rate for a successful landing.
- Most of the launch sites are located near coastlines with critical infrastructure in proximity.
- Most of the payload is between the range 2000 and 4000 kg with FT 1 booster providing the most successful launches.
- From the given features of launch site, booster version and other factors, we can determine with 83% accuracy.

Thank you!