Winning Space Race with Data Science

Jennifer Milano
30 July 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection using REST API and Web Scraping
  - Exploratory Data Analysis (EDA) with Data Visualization (Matplotlib and Pandas)
  - EDA with SQL
  - Interactive Map with Folium
  - Interactive Dashboards with Plotly Dash
  - Predictive Analysis using Machine Learning
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive maps and dashboard
  - Predictive results

Introduction

- Project background and context
  - SpaceX launches Falcon 9 rockets with a lower cost per launch at about $62m as compared to other companies (typically about $165m). This can be credited to the fact that SpaceX can land and re-use the first stage of the rocket.
  - If predictions can be made about whether the first stage will land or not, then we can determine the cost of a launch. We can then use this information to assess if an alternate company should bid against SpaceX for a rocket launch.

- Problems you want to find answers
  - The aim of this project is to successfully predict if the SpaceX Flacon 9 will land successfully.

Section 1

Methodology

Methodology – Executive Summary

I.Data Collection
- •Using GET requests to the SpaceX REST API
- •Web Scraping

II.Data Wrangling
- •Using .fillna() method to remove NaN values
- •Using .value_counts() method to determine:
  - •Number of launches on each site
  - •Number of occurrences of each orbit
  - •Number and occurrence of mission outcome per orbit type
- •Creating a landing outcome label that shows the following:
  - •0 when the booster did not land successfully
  - •1 when the booster landed successfully

III.Exploratory Data Analysis
- •Using SQL queries to manipulate and evaluate the SpaceX dataset
- •Using Pandas and Matplotlib to visualize relationships between variables and determine patterns

IV.Interactive Visual Analytics
- •Geospatial analytics using Folium
- •Creating an interactive dashboard using Plotly Dash

V.Data Modeling and Evaluation
- •Using Scikit-Learn to:
  - •Pre-process (standardize and normalize the data)
  - •Split the data into training and testing data using .train_test_split()
  - •Train different classification models
  - •Find hyperparameters using .GridSearchCV()
- •Plotting confusion matrices for each classification model
- •Assessing the accuracy of each classification model

Data Collection – SPACEX REST API

Data Collection performed using the SpaceX REST API to retrieve data about launches.
This process includes gathering information on the types of rocket used, payload delivered, launch specification, landing specification, and landing outcome for each entry.

STEP 1:

- Make a GET response to the SpaceX REST API
- Convert it to a .json file then into a Pandas DataFrame

STEP 2:

- Use custom logic to clean the data
- Define lists for the data to be stored in
- Call custom functions to retrieve data and fill the lists
- Use list values as a dictionary and construct the dataset

STEP 3:

- Create Pandas DataFrame from the constructed dictionary dataset

STEP 4:

- Filter the DataFrame to only include Falcon 9 launches
- Reset the FlightNumber column
- Replace missing values of PayloadMass with mean PayloadMass

Data Collection – Web Scraping

Data Collection using web scraping done with BeautifulSoup to collect Falcon 9 historical launch records from a Wikipedia page titled Falcon 9 and Falcon Heavy launches.

STEP 1:
- Request the HTML page from the static URL

STEP 2:
- Create a BeautifulSoup object
- Find all tables within the HTML page

STEP 3:
- Collect all column header names from the tables found within the HTML page

STEP 4:
- Use the column names as keys in a dictionary
- Use custom functions and logic to parse all launch tables to fill dictionary values

STEP 5:
- Convert the dictionary to Pandas DataFrame ready for export

Data Wrangling – Using Pandas

- The SpaceX dataset contains various launch facilities. We can identify the launch facilities in the LaunchSite column.
- Each launch aims toward a dedicated orbit. The orbit types can be found in the Orbit column.
- Using the .value_counts() method, we can determine the following information:
  - Number of launches for each site
  - Number and occurrence of each orbit
  - Number and occurrence of landing outcome per orbit type
- The landing outcome is shown in the Outcome column.
- To determine whether a booster will successfully land, it is best to have a binary column (i.e., where the value is 1 or 0, representing the landing status).
- The is done by:

Step 1: Defining a set of unsuccessful outcomes, bad_outcome.

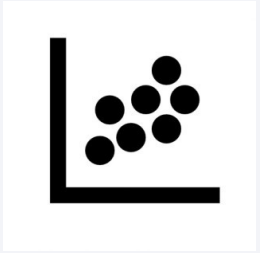
Step 2: Creating a list, landing_class, where the element is 0 if the corresponding row in Outcome is in the set bad_outcome, otherwise, it's 1.

Step 3: Create a Class column that contains the values from the list landing_class.

Step 4: Export the DataFrame as a .csv file.

Exploratory Data Analysis with Data Visualization

- Scatter Plots
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload vs. Launch Site
  - Orbit vs. Flight Number
  - Payload vs. Orbit Type
  - Orbit vs. Payload Mass



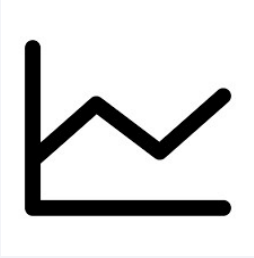
Scatter plots show relationships and/or correlation between variables.

- Bar Charts
  - Success rate vs. Orbit

Bar charts show the relationship between numeric and categoric variables.

- Line Chart
  - Success rate vs. Year

Line charts show data variables and their trends. Line graphs can help to show global behavior and make prediction for unseen data.



Exploratory Data Analysis with SQL

- We performed SQL queries to gather and understand data from dataset:
  - Displaying 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 records which will display the month names, faiilure landing_ouutcomes in drone ship, booster versions, launch_site for the months in year 2015.
  - Rank the count of successful landiing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Build an Interactive Map with Folium

The following steps were taken to visualize the launch data on an interactive map:

1. Mark all launch sites on a map
   - Initialize the map using a Folium Map object
   - Add a folium.Circle and folium.Marker for each launch site on the launch map
2. Mark the success/failed launches for each site on a map
   - As many launches have the same coordinates, it makes sense to cluster them together.
   - Before clustering them, assign a marker color of successful (class = 1) as green, and failed (class = 0) as red. To put the launches into clusters, for each launch, add a folium.Marker to the MarkerCluster() object.
   - Create an icon as a text label, assigning the icon_color as the marker_color determined previously.
3. Calculate the distances between a launch site to its proximities
   - To explore the proximities of launch sites, calculations of distances between points can be made using the Lat and Long values.
   - After marking a point using the Lat and Long values, create a folium.Marker object to show the distance.
   - To display the distance line between two points, draw a folium.PolyLine and add this to the map.

Build a Dashboard with Plotly Dash

The following plots were added to a Plotly Dash dashboard to have an interactive visualization of the data:

- Pie chart showing the total successful launches per site
  - This makes it clear to see which sites are most successful
  - The chart could also be to see the success/failure ratio for an individual site

- Scatter graph to show the correlation between outcome (success or not) and payload mass (kg)
  - This could be filtered (using a RangeSlider() object) by ranges of payload masses
  - It could also be filtered by booster version

Predictive Analysis (Classification)

MODEL DEVELOPMENT

- To prepare the dataset for model development:
  - Load dataset
  - Perform necessary data transformations (standardize and pre-process)
  - Split data into training and test data sets
  - Decide which type of machine learning algorithms are most appropriate
- For each chosen algorithm:
  - Create a GridSearchCV object and a dictionary of parameters
  - Fit the object to the parameters
  - Use the training data set to train the model

MODEL EVALUATION

- For each chosen algorithm:
  - Using the output GridSearchCV object:
  - Check the tuned hyperparameters (best_params_)
  - Check the accuracy (score and best_score_)
  - Plot and examine the Confusion Matrix

FINDING THE BEST MODEL

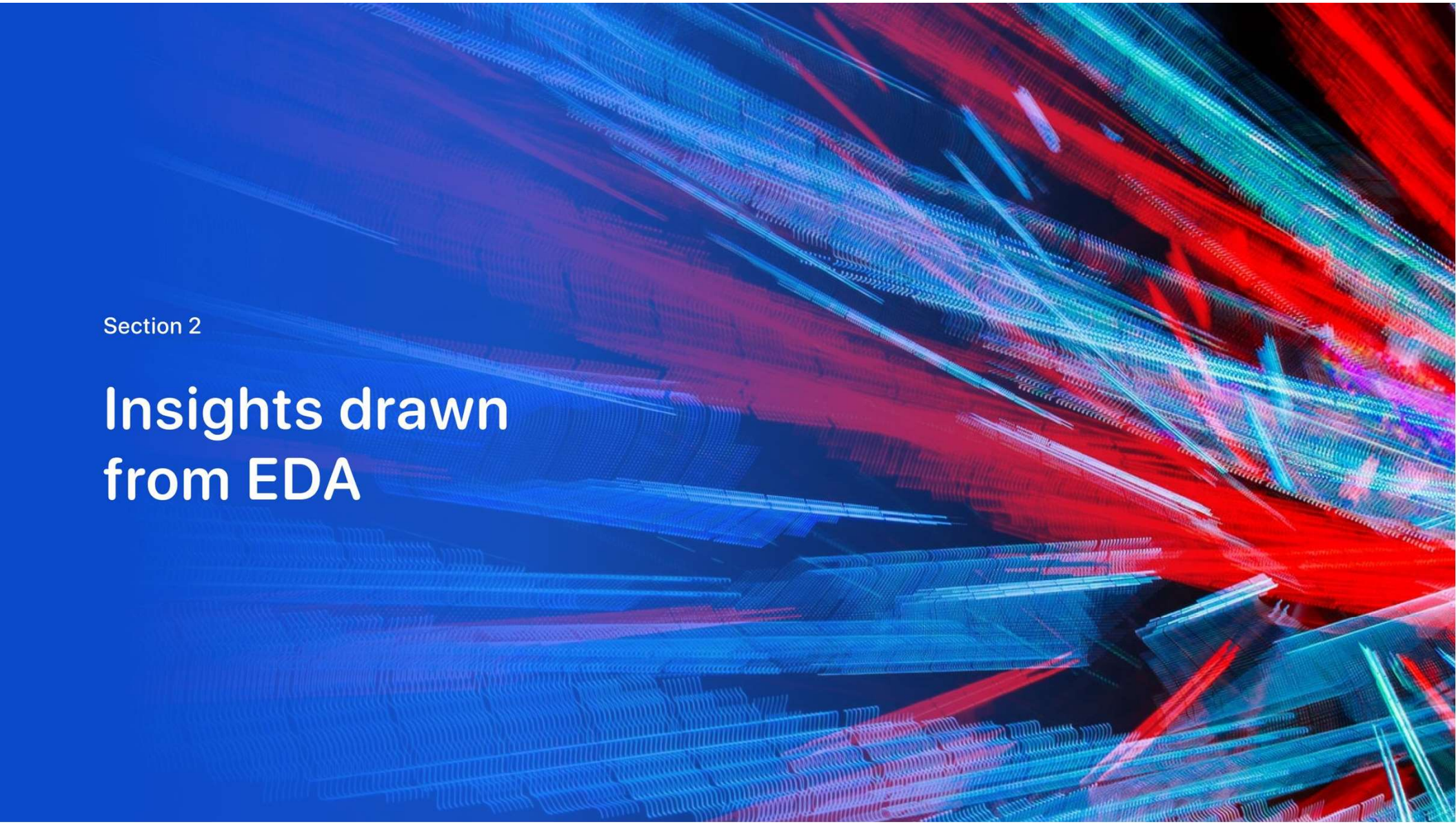
- Review the accuracy scores for each algorithm
- The model with the highest accuracy score is determined as the best performing model

Results

- Exploratory data analysis results
  - There is a relationship between the payload, the orbits and the success of the landing.
  - SSO orbit has a 100% success rate for a payload below 6000kg.
  - The launch success rate gets better with time.

- Interactive Analysis
  - Most of the launch sites are close to the equator and are in close proximity to the coastline but also far from city centers while being accessible by highways and railways.

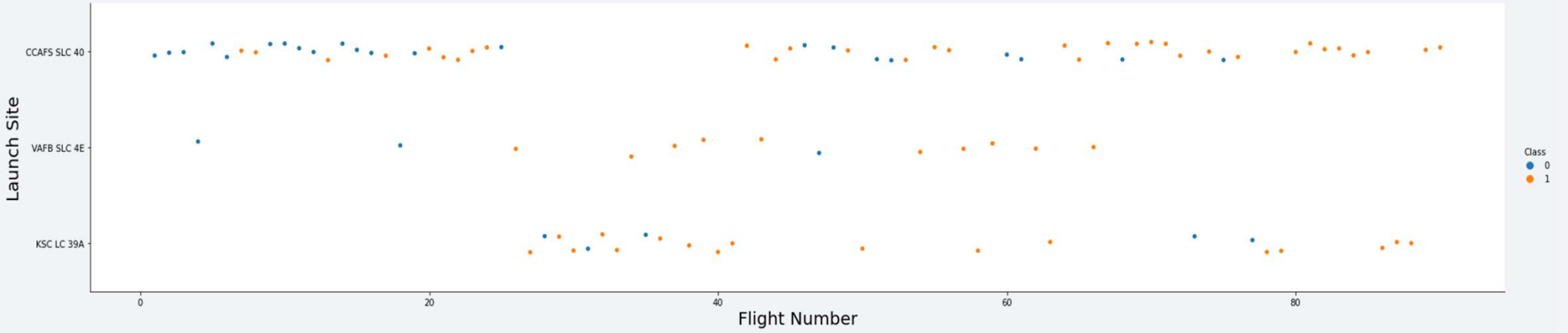
- Predictive Analysis
  - Among all the prediction methods: it is the decision tree that slightly performs better.



Section 2

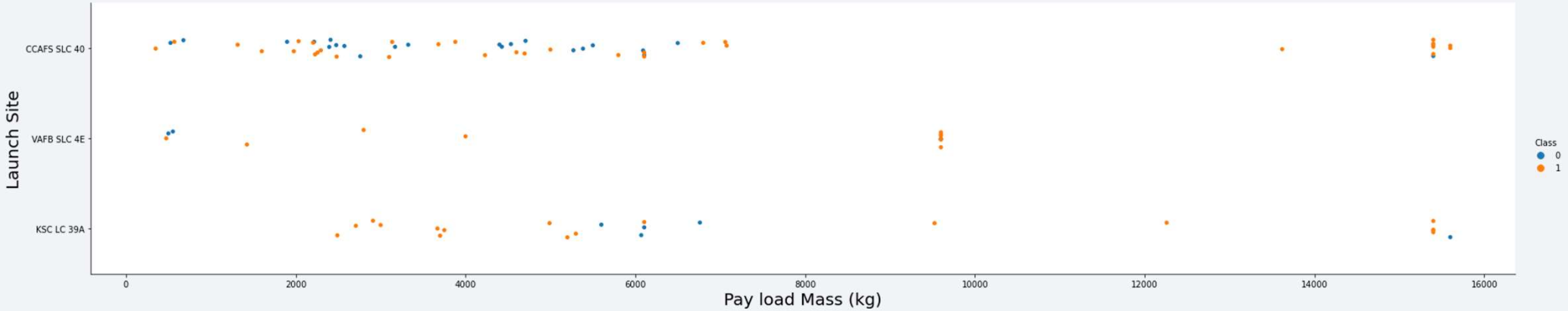
Insights drawn from EDA

Flight Number vs. Launch Site



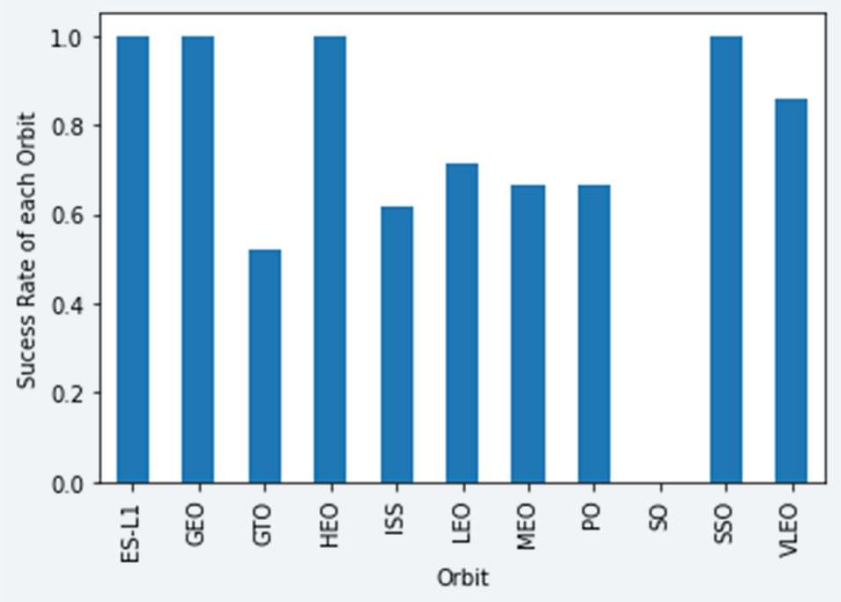
The success of a landing is correlated to the number of flights. The more launches are done, the more landings are successful.

Payload vs. Launch Site



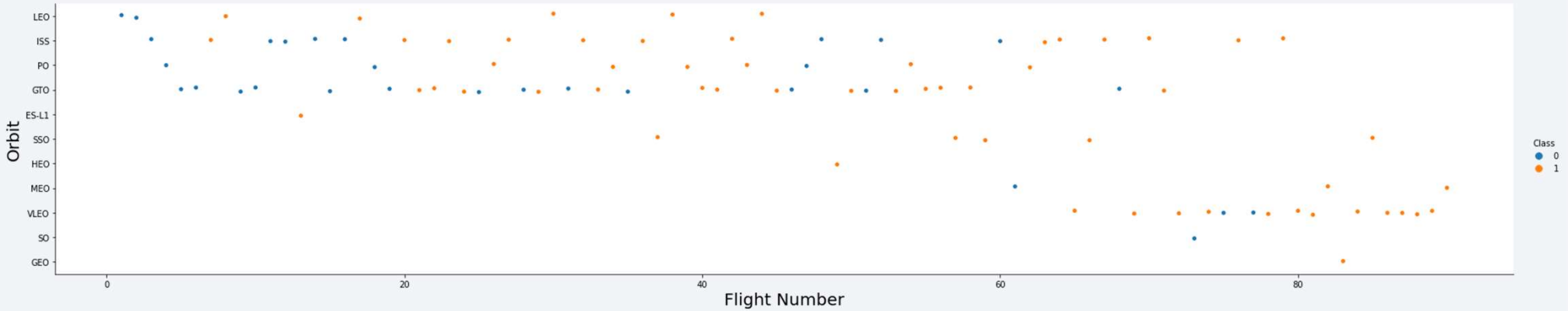
- Above a payload mass of around 7000 kg, there are very few unsuccessful landings,but there is also far less data for these heavier launches.
- There is no clear correlation between payload mass and success rate for a given launch site.
- All sites launched a variety of payload masses, with most of the launches from CCAFS SLC 40 being comparatively lighter payloads (with some outliers).

Success Rate vs. Orbit Type



With this plot, we can see success rate for different orbit types. We note that ES-L1, GEO, HEO, SSO have the best success rate.

Flight Number vs. Orbit Type

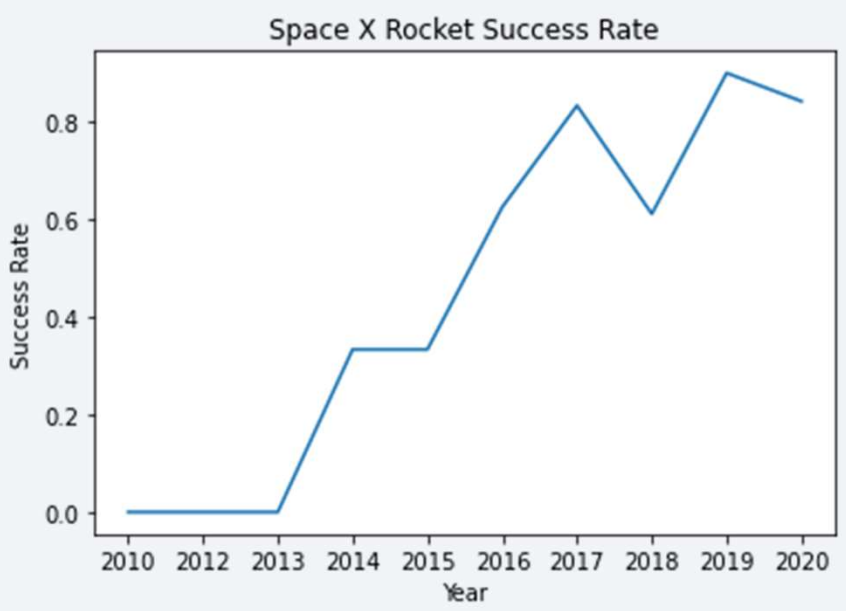


- The 100% success rate of GEO, HEO, and ES-L1 orbits can be explained by only having 1 flight into the respective orbits.
- The 100% success rate in SSO is more impressive, with 5 successful flights.
- There is little relationship between Flight Number and Success Rate for GTO.
- Generally, as Flight Number increases, the success rate increases. This is most extreme for LEO, where unsuccessful landings only occurred for the low flight numbers (early launches).

Payload vs. Orbit Type

The weight of the payloads can have a great influence on the success rate of the launches in certain orbits. For example, heavier payloads improve the success rate for the LEO orbit. Another finding is that decreasing the payload weight for a GTO orbit improves the success of a launch.

Launch Success Yearly Trend



Since 2013, we can see an increase in the Space X Rocket success rate.
Between 2010 and 2013, all landings were unsuccessful.

All Launch Site Names

In [9]:
```
%sql select distinct launch_site from spacextbl;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[9]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

In [10]:
```
%sql select * from spacextbl \
     where launch_site like 'CCA%' limit 5
```

```
* sqlite:///my_data1.db
Done.
```

Out[10]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Lan
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Fai
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Fai
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	

Total Payload Mass

In [11]:
```
%sql select sum(payload_mass__kg_) from spacextbl \
     where customer = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[11]:

sum(payload_mass__kg_)
45596.0

Average Payload Mass by F9 v1.1

In [12]:
```
%sql select avg(payload_mass__kg_) from spacextbl where booster_version = 'F9 v1.1'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[12]:

avg(payload_mass__kg_)
2928.4

First Successful Ground Landing Date

```
[16]: %sql select min(date) from SPACEXTBL where landing_outcome = "Success (ground pad)"
```

```
 * sqlite:///my_data1.db
Done.
```

[16]:

min(date)
01/08/2018

The first successful ground landing was on 01 August 2018.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
[15]: %sql select booster_version from SPACEXTBL where landing_outcome = 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
 * sqlite:///my_data1.db
Done.
```

[15]:

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

The best boosters for drone ship landing with a payload mass between 4000 kg and 6000 kg are:
F9 FT B1022
F9 B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

In [16]:
```
%sql select mission_outcome, count (mission_outcome) from spacextbl group by mission_outcome
```

```
 * sqlite:///my_data1.db
Done.
```

Out[16]:

Mission_Outcome	count (mission_outcome)
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

There were a total of 99 successful missions.

Boosters Carried Maximum Payload

```
[17]: %sql select booster_version, payload_mass__kg_ from spacextbl where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl)
```

```
 * sqlite:///my_data1.db
Done.
```

[17]:

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

[18]:
```
%sql select *,  substr(Date, 4, 2) as 'month names' from spacextbl where landing_outcome = 'Failure (drone ship)' and substr(Date,7,4)='2015'
```

```
 * sqlite:///my_data1.db
Done.
```

[18]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome	month names
01/10/2015	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395.0	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)	10
14/04/2015	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898.0	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)	04

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

[125]:
```
%%sql select Landing_Outcome, count(Landing_Outcome) as count
from spacextbl
where Date between '04-06-2010' and '20-03-2017'
and (Landing_Outcome like 'Success%')
group by Landing_Outcome
order by count desc
```

```
 * sqlite:///my_data1.db
Done.
```

[125]:

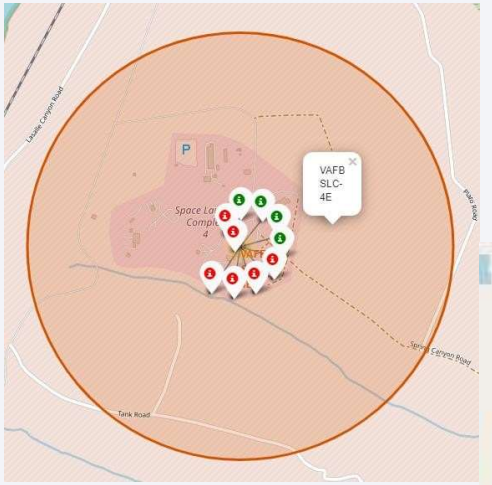
Landing_Outcome	count
Success	20
Success (drone ship)	8
Success (ground pad)	7

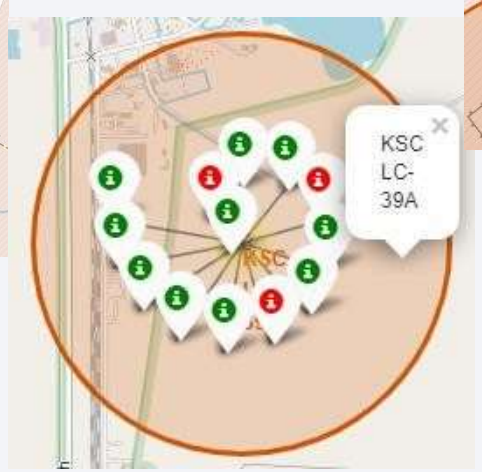

Section 4
Launch Sites
Proximities Analysis

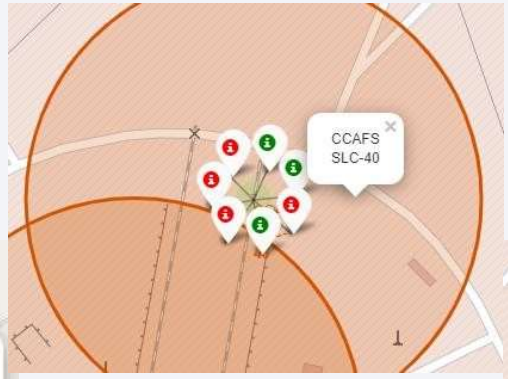
Launch Sites Locations

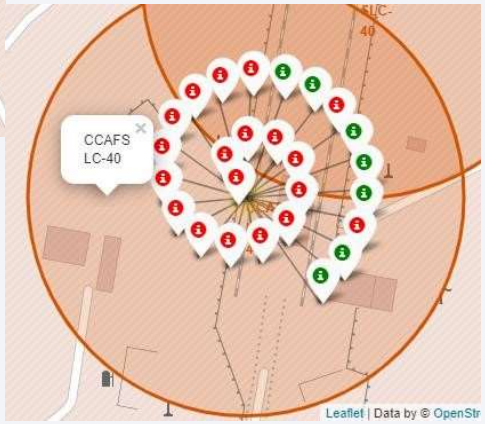
We see that Space X launch sites are located on the coast of the United States and close to the equator.

Failed and Successful Launches by Site



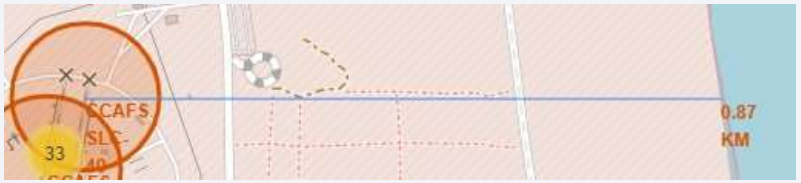


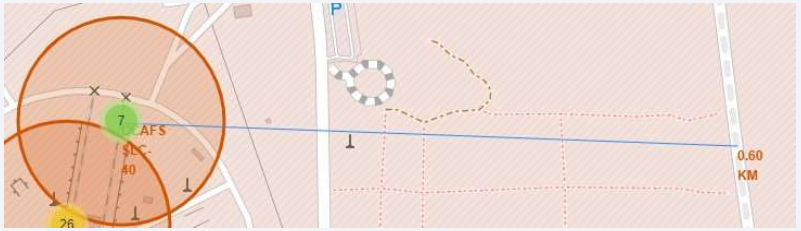


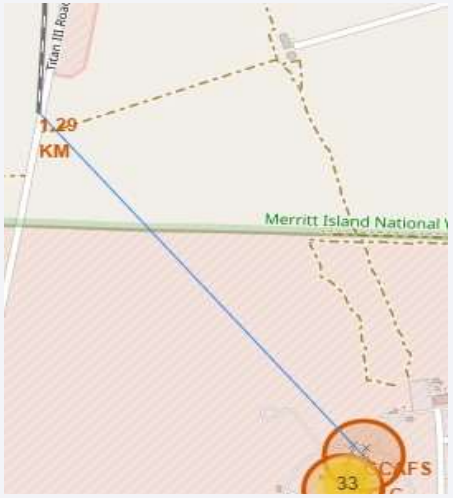


Green marker represents successful launches. Red marker represents unsuccessful launches. We note that KSCLC-39A has a higher launch success rate.

Proximity of Launch Site to Points of Interest







Using CCAFS SLC-40 as an example site, the following questions can be asked to understand more about the placement of launch sites:

Are launch sites in close proximity to railways?

- Yes, the coastline is only 0.87 km due East.

Are launch sites in close proximity to highways?

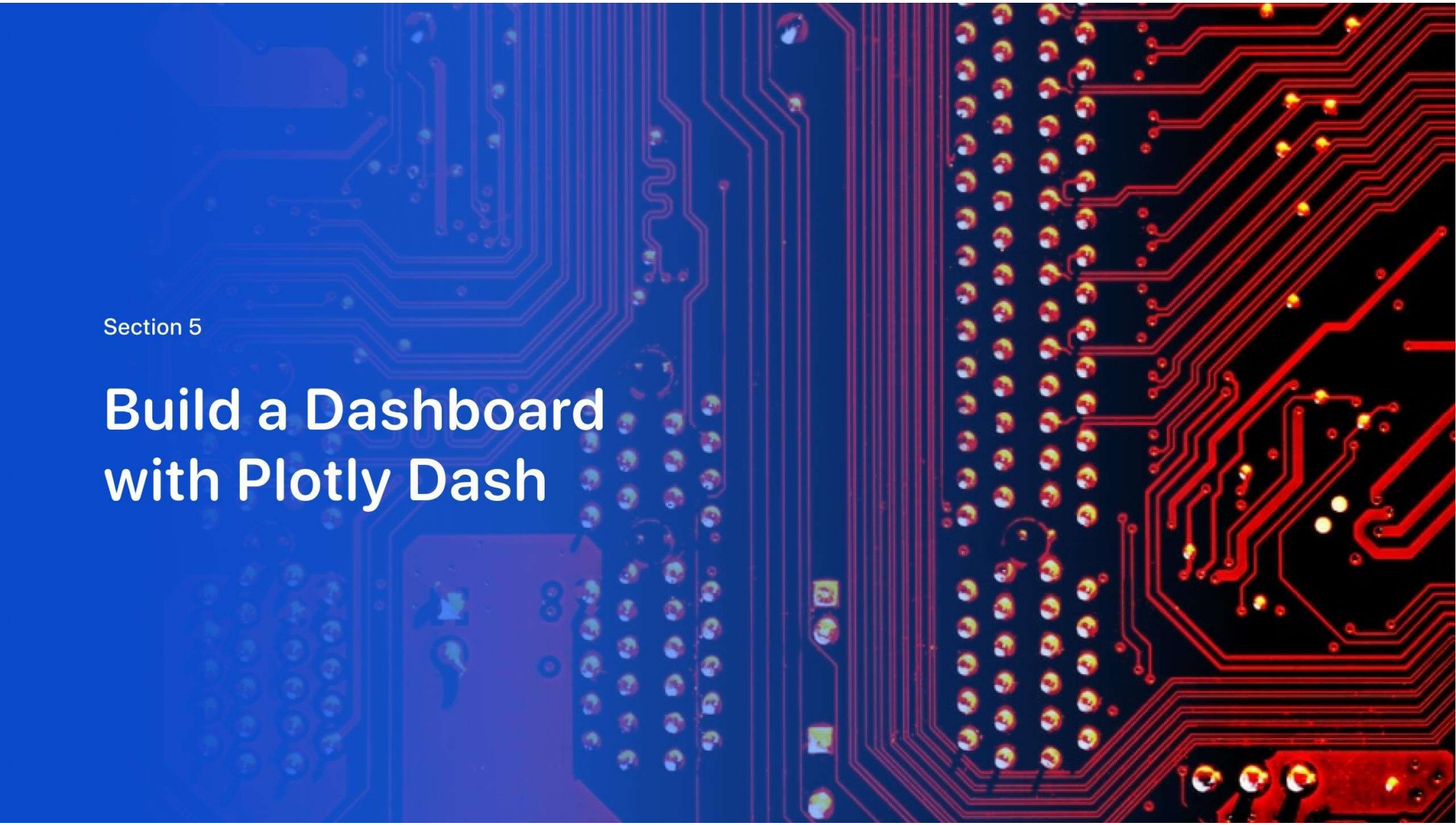
- Yes, the nearest highway is only 0.59 km away.

Are launch sites in close proximity to railways?

- Yes, the nearest railway is only 1.29 km away.

Do launch sites keep a certain distance away from cities?

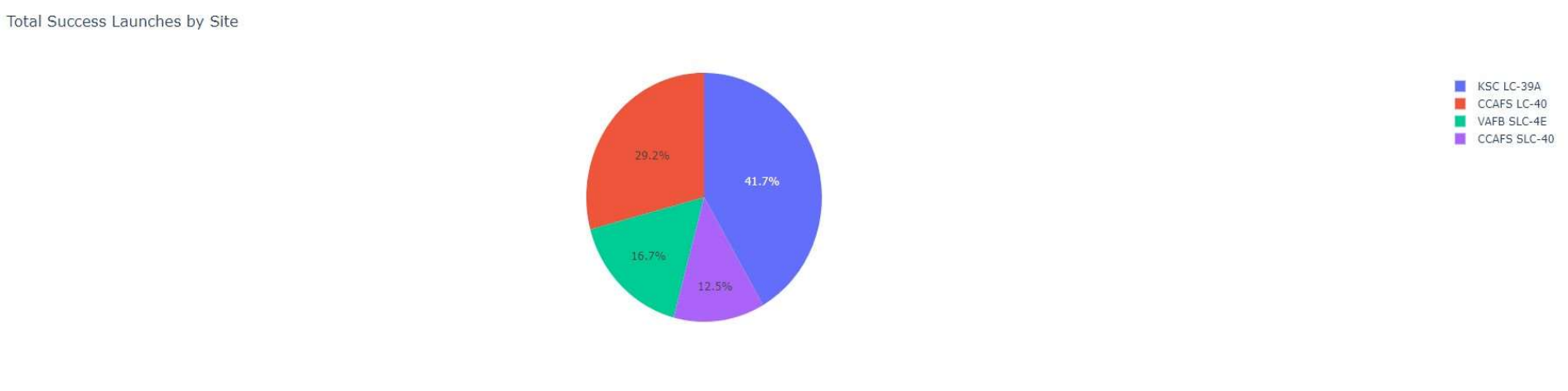
- Yes, the nearest city is 51.74 km away.



Section 5

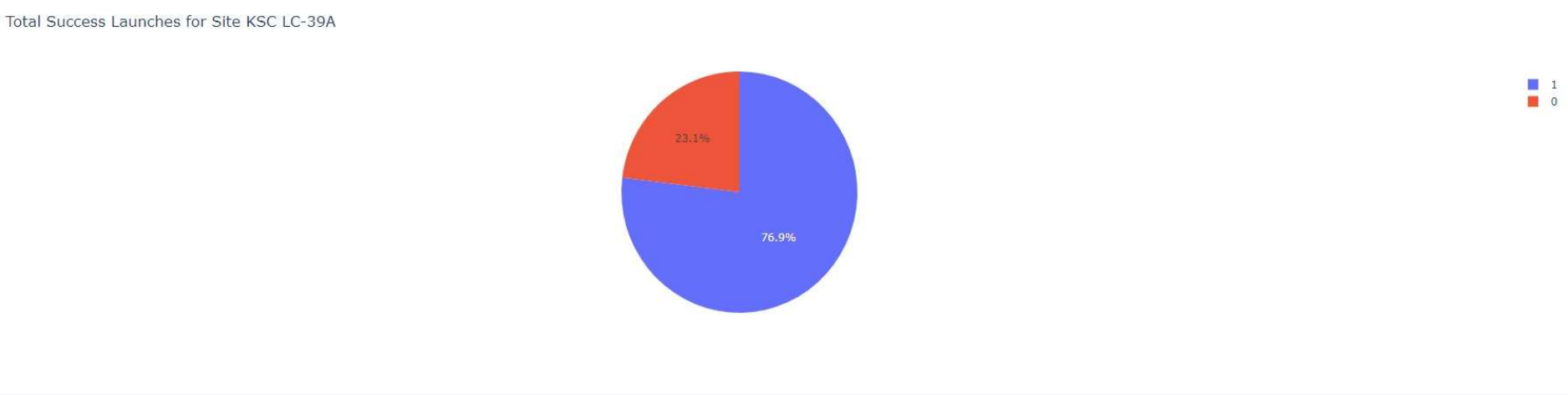
Build a Dashboard with Plotly Dash

Launch Success Count for All Sites



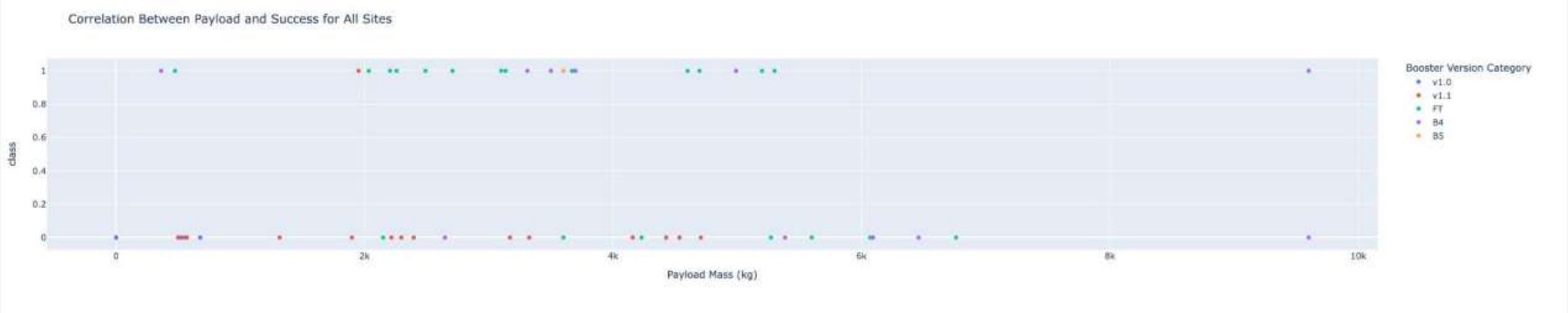
KSC LC-39A represents 41.7% of the total successful launches.

Total success launches for Site KSC LC-39A



KSC LC-39A has achieved a 76.9% success rate while getting a 23.1% failure rate.

Payload mass vs Outcome for all sites



Low weighted payloads have a better success rate than the heavy weighted payloads.

Section 6

Predictive Analysis (Classification)

Classification Accuracy

In [26]:
```
lr_score = logreg_cv.score(X_test, Y_test)
print("score :", lr_score)
```

```
score : 0.8333333333333334
```

In [33]:
```
svm_cv_score = svm_cv.score(X_test, Y_test)
print('score :',svm_cv_score)
```

```
score : 0.8333333333333334
```

In [46]:
```
tree_cv_score = svm_cv.score(X_test, Y_test)
print("score :",tree_cv_score)
```

```
score : 0.8333333333333334
```

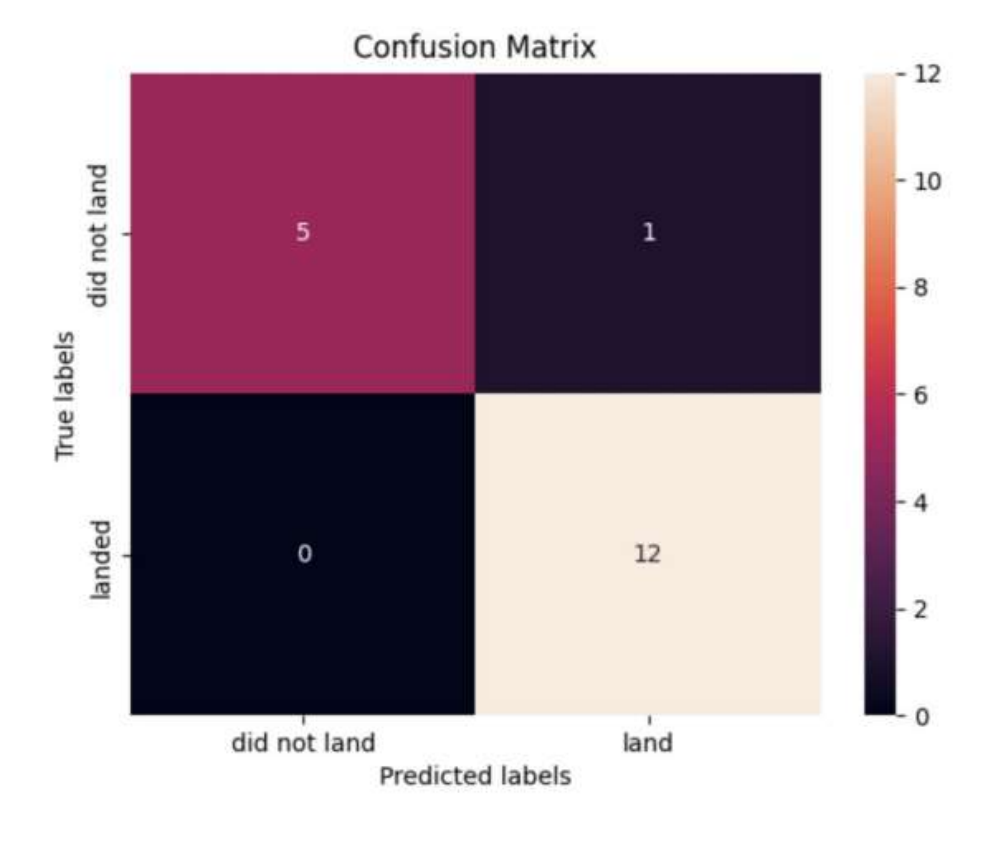
In [54]:
```
knn_cv_score = knn_cv.score(X_test, Y_test)
print('knn score: ', knn_cv_score)
```

```
knn score:  0.8333333333333334
```

Logistic regression, support vector machine, decision tree, and K nearest neighbors have the same accuracy: 0.833.

Confusion Matrix

- As shown previously, best performing classification model is the K-Nearest Neighbors model, with an accuracy of 94.44%.
- This is explained by the confusion matrix, which shows only 1 out of 18 total results classified incorrectly (a false positive, shown in the top-right corner).
- The other 17 results are correctly classified (5 did not land, 12 did land).



Conclusions

As the number of flights increases, the rate of success at a launch site increases, with most early flights being unsuccessful (i.e. with more experience, the success rate increases.)

- Between 2010 and 2013, all landings were unsuccessful (as the success rate is 0).
- After 2013, the success rate generally increased, despite small dips in 2018 and 2020.
- After 2016, there was always a greater than 50% chance of success.

Orbit types ES-L1, GEO, HEO, and SSO, have the highest (100%) success rate.

- The 100% success rate of GEO, HEO, and ES-L1 orbits can be explained by only having 1 flight into the respective orbits.
- The 100% success rate in SSO is more impressive, with 5 successful flights.
- The orbit types PO, ISS, and LEO, have more success with heavy payloads:
  - VLEO (Very Low Earth Orbit) launches are associated with heavier payloads, which makes intuitive sense.

The launch site KSC LC-39 A had the most successful launches, with 41.7% of the total successful launches, and also the highest rate of successful launches, with a 76.9% success rate.

The success for massive payloads (over 4000 kg) is lower than that for low payloads (this means a smaller payload translates to a higher success rate).

The best performing classification model is the K-Nearest Neighbors model, with an accuracy of 94.44%.

Thank you!