Winning Space Race with Data Science

Steven Olawale

18th May 2023

Outline

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

Executive Summary

- Summary of methodologies

  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction

- Summary of all results

  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  Space X 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 Space X can reuse the first stage. This information can be used if an alternate company wants to bid against space X for a rocket launch. The goal of the project is to create a machine learning pipeline if the first stage will land successfully.

- Problems you want to find answers

  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing
  - What operating conditions needs to be in place to ensure a successful landing program

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Space X API (https://api.spacexdata.com/v4/rockets/)
  - WebScraping (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- Perform data wrangling
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash

Methodology

Executive Summary

- Perform predictive analysis using classification models
  - Data that was collected until this step were normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combination of parameters

Data Collection

- Describe how data sets were collected.
  - Datasets were collected from SpaceX API (https://api.spacexdata.com/v4/rockets/) and from Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches), using web scraping technics.

Data Collection – SpaceX API

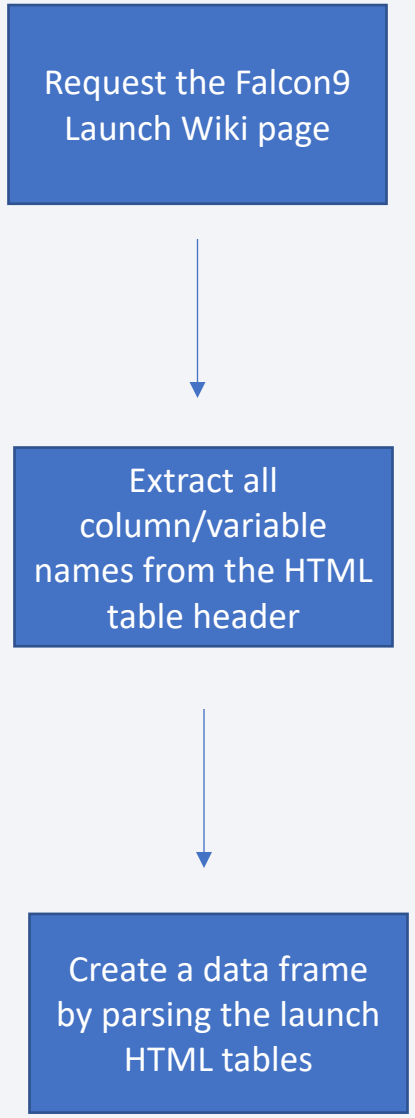
- SpaceX offers a public API where data can be obtained before using

- Github URL: https://github.com/stevedeine/IBM-Capstone/blob/main/Data%20Collection.ipynb

Data Collection - Scraping

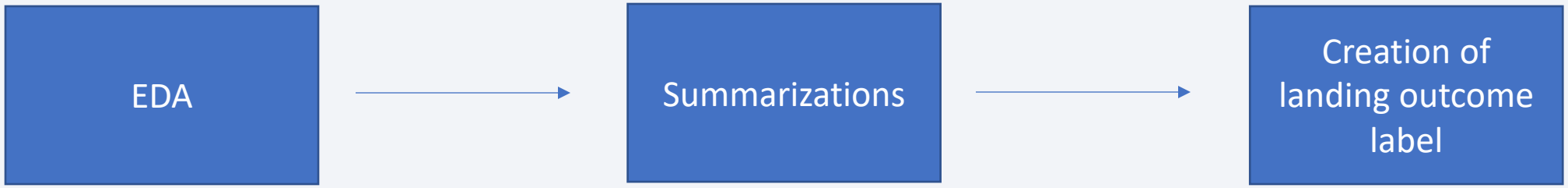
- Data from Space X launches can also be obtained from Wikipedia

- Data are downloaded from Wikipedia according to the flowchart and then persisted
- Github URL: https://github.com/stevedeine/IBM-Capstone/commit/c3fe0458cc90725516124d2f552ebbd7f8df3f38



Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset
- Then the summary launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated
- Finally, the landing outcome label was created from Outcome column



- Github URL: https://github.com/stevedeine/IBM-Capstone/blob/main/Data%20Wrangling.ipynb

EDA with Data Visualization

- The following SQL queries were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begins with the string 'CCA';
  - Total pay load mass carried by boosters launched by NASA (CRS);
  - Average payload mass carried by booster version f9 v1.1;
  - Date when the first successful landing outcome in ground pad was achieved;
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000kg;
  - Total number of successful and failed mission outcomes
  - Names of the booster versions which have carries the maximum payload mass;
  - Failed landing outcomes in droneship, their booster versions, and launch site names for in year 2015; and
  - Rank of the count of launding outcomes (such as Failure (droneship) r Success (ground pad) between the date 2010-06-04 and 2017-03-20
- Github URL - https://github.com/stevedeine/IBM-Capstone/blob/2d0ef393eed8ac3d431de8b5c1c7317ca50d14ef/EDA%20with%20visualisation.ipynb

EDA with SQL

- To explore data, scatterplots and bar plots were used to visualize the relationship between pair of features:
- Payload Mass vs Flight Number, Launch Site vs Flight Number, Launch Site vs Payload Mass, Orbit vs Flight Number, Payload vs Orbit
- Visualization
- GitHub URL - https://github.com/stevedeine/IBM-Capstone/blob/3245ccbc31fd0c4a69bd047aeeb948ab14c16799/jupyter-labs-eda-sql-coursera-3-3.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
- Markers indicate points like launch sites;
- Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
- Marker clusters indicates group of events in each coordinate, like launches in a launch site; and
- Lines are used to indicate distances between two coordinates
- GitHub URL - https://github.com/stevedeine/IBM-Capstone/blob/3ea01aa2e2a17972ff76496f9be024d42ef7a679/IBM-DS0321EN-SkillsNetwork_labs_module_3_lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- An interactive dashboard was built with plotly dash
- We plotted pie charts showing the total launches by a certain site
- We plotted scatter graph showing the relationship between Outcome and Payload Mass(kg) for the different booster version.
- GitHub link - https://github.com/stevedeine/IBM-Capstone/commit/a8ba542ec868634fb11219faad7f3cf5c1a96391

Predictive Analysis (Classification)

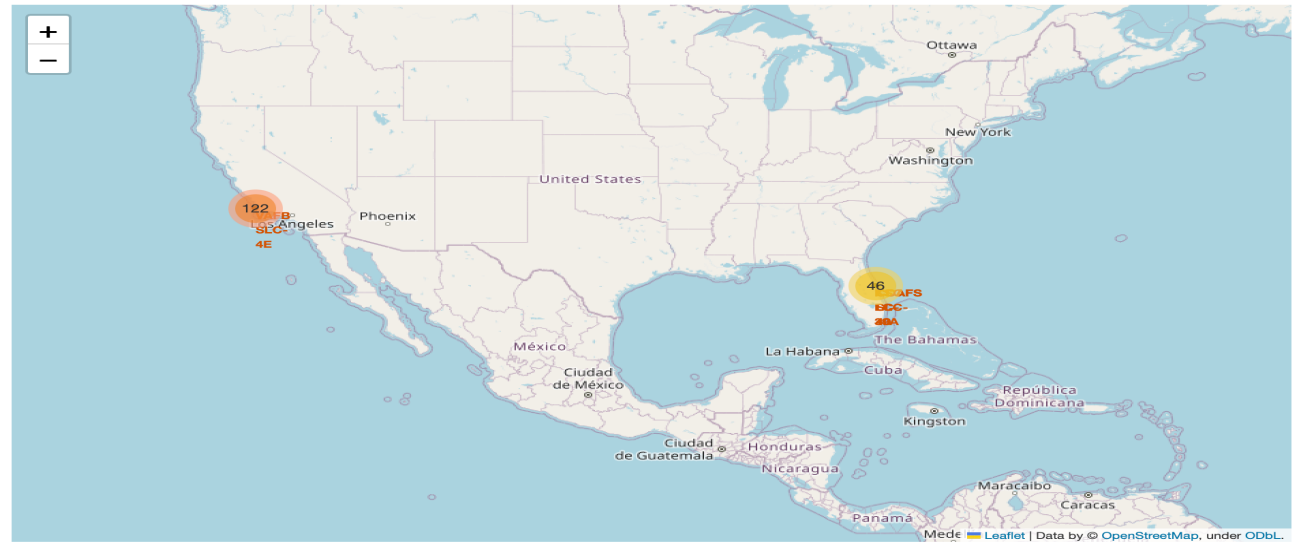
- The data was loaded using numpy and pandas, transformed the data, split our data into training and testing
- We built different machine learning models and tune different hyperparameters using GridsearchCV
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning
- We found the best performing classification model.
- GitHub link - https://github.com/stevedeine/IBM-Capstone/blob/2d0ef393eed8ac3d431de8b5c1c7317ca50d14ef/IBM-DS0321EN-SkillsNetwork_labs_module_4_SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

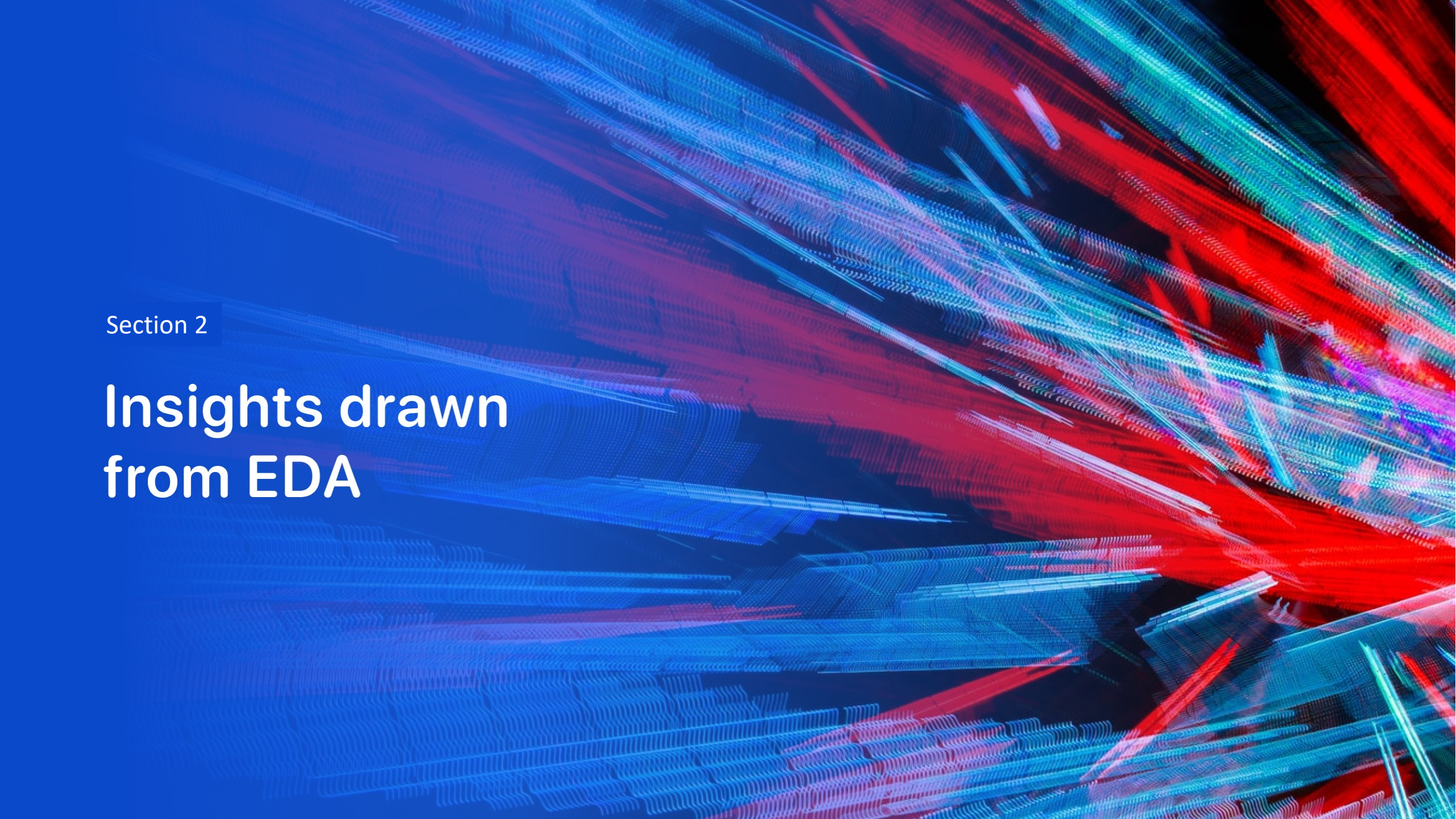
Results

- Exploratory data analysis results
  - Space X uses 4 different launch sites;
  - The first launches were done to Space X itself and NASA
  - The average payload of F9 v1.1 booster is 2928 kg;
  - The first success landing outcome happened in 2015, 5 years after the first launch;
  - Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
  - Almost 100% of mission outcomes were successful;
  - Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015

Results

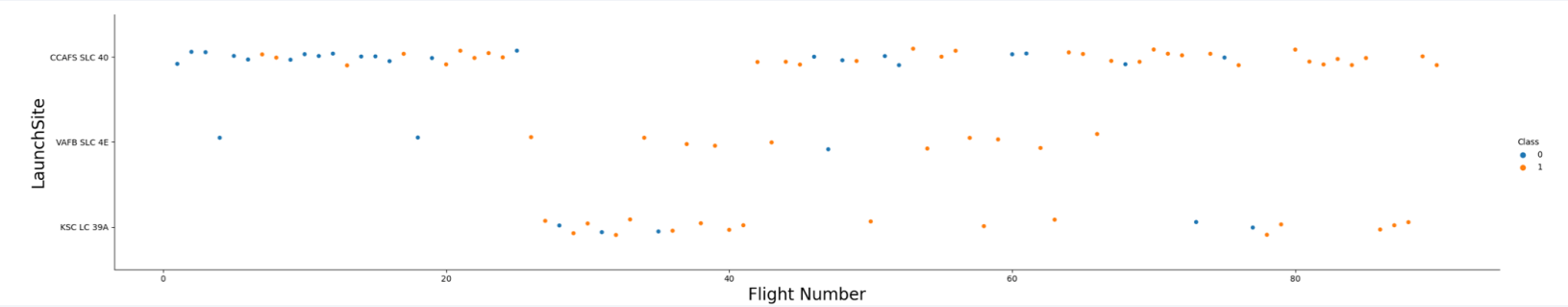
- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east coast launch sites




Section 2
Insights drawn from EDA

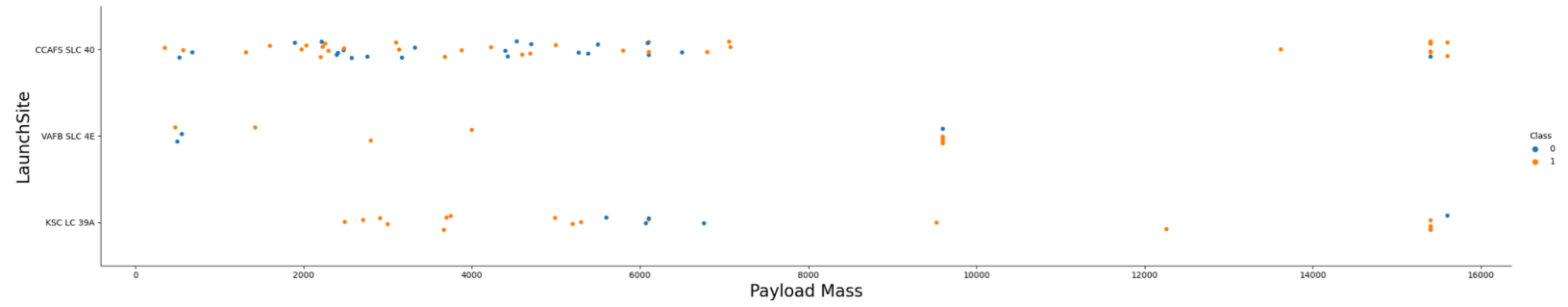
Flight Number vs. Launch Site

- From the plot, I was able to deduce that the larger the flight amount at a launch site, the greater the success rate at a launch site



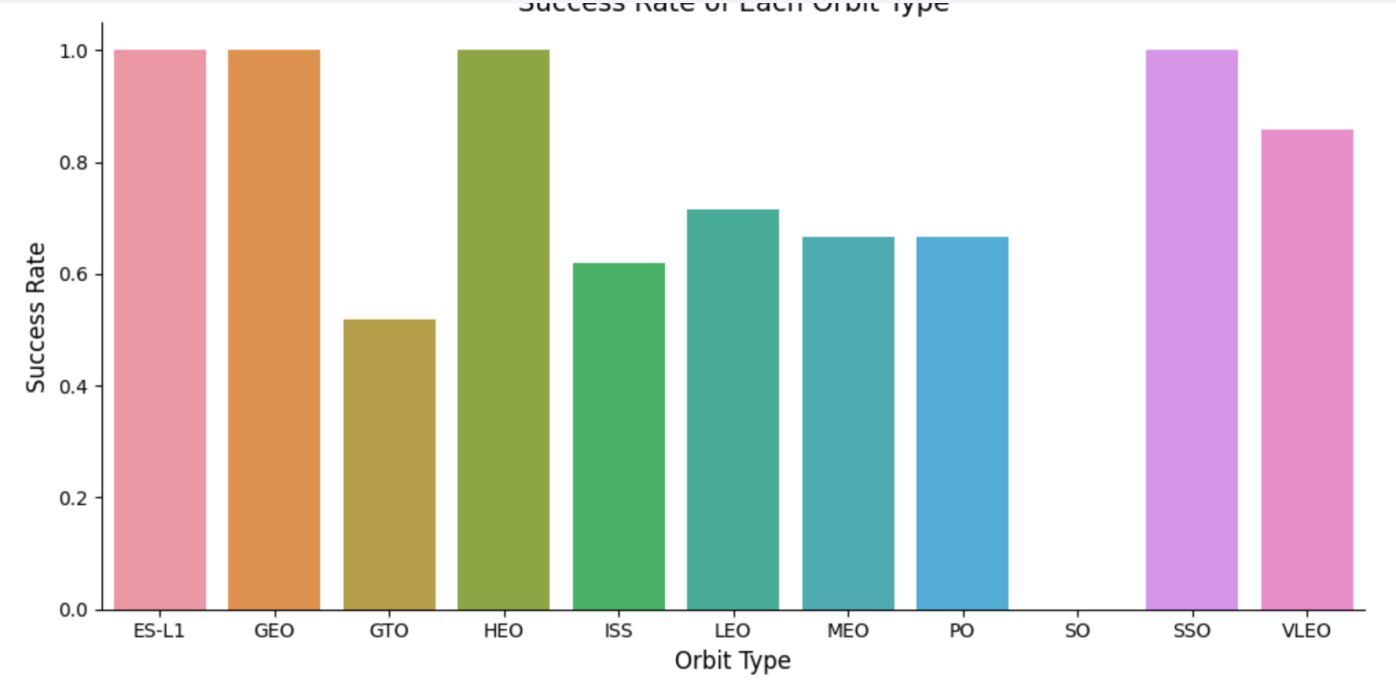
Payload vs. Launch Site

- From the plot, I was able to deduce that there was very little success in launch site KSC LC 39A and VAFB SLC 4E and more success in CCAFS SLC 40



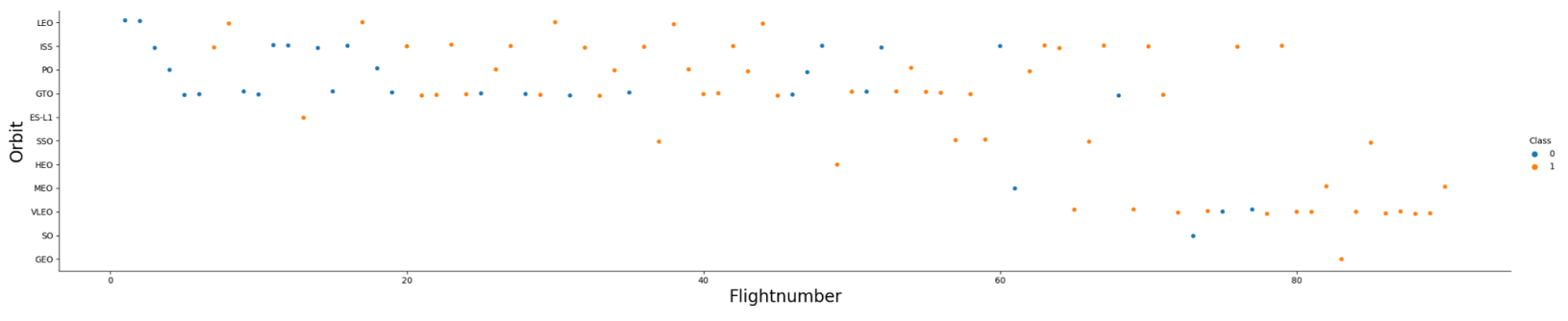
Success Rate vs. Orbit Type

- From the plot, we can see that ES-I1, GEO, HEO, SSO had the most success rate followed by VLEO and LFO



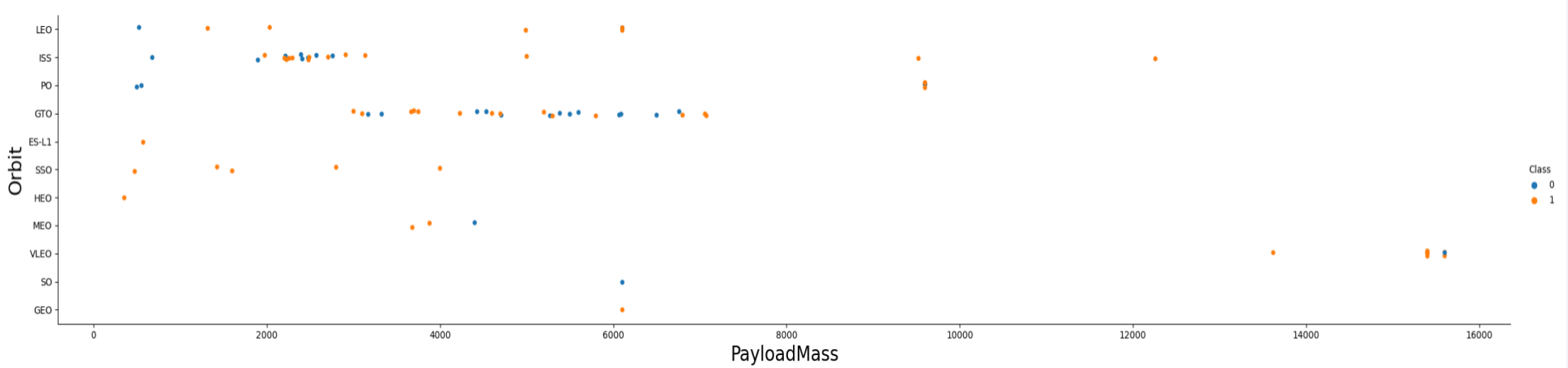
Flight Number vs. Orbit Type

- The plot below shows the Flight Number vs. Orbit type. We observe that there was success rate overtime in the VLEO orbit



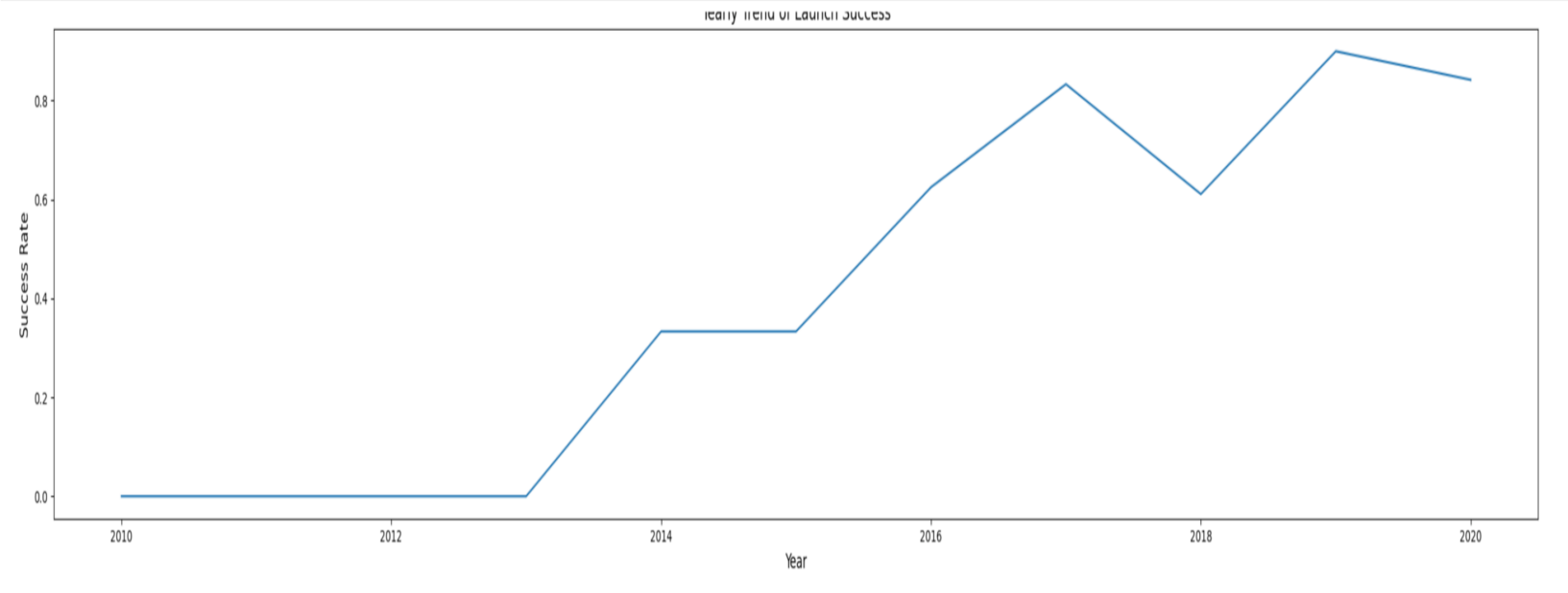
Payload vs. Orbit Type

- There is no relationship between payload and success rate to orbit GTO, ISS orbit has the widest range of payload and a good success rate with very little launches in SO and GEO.



Launch Success Yearly Trend

- Success rate started increasing from 2013 through till 2020



All Launch Site Names

- There are 4 launch sites available in the dataset

```
In [5]: %sql select distinct(launch_site) from spacex
 * ibm_db_sa://vry32322:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
Out[5]:
```

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

Launch Site Names Begin with 'CCA'

- We can that 5 records where launch sites begin with `CCA'

```
In [6]: %sql select * from spacex where launch_site like 'CCA%' limit 5
 * ibm_db_sa://vry32322:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

Out[6]:

DATE	time_utc	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcom
010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failur (parachute
010-12-08	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	Failur (parachute
012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attemp
012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attemp
013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

Total Payload Mass

- Total Payload carried by booster = 48213kg

```
In [7]: %sql select SUM(Payload_mass_kg_) from spacex where customer like '%NASA (CRS)%'
 * ibm_db_sa://vry32322:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
Out[7]:
1
48213
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2534

Display average payload mass carried by booster version F9 v1.1

```
[9]: %sql select AVG(Payload_mass_kg_) from spacex where booster_version like '%F9 v1.1%'
```

```
 * ibm_db_sa://vry32322:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

[9]:

1
2534

First Successful Ground Landing Date

- The date of the first launch 2015

```
In [10]: %sql select min(date) from spacex where landing_outcome = 'Success (ground pad)'
 * ibm_db_sa://vry32322:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
Out[10]:
1
2015-12-22
```

Successful Drone Ship Landing with Payload between 4000 and 6000

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

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
In [14]: %sql select Booster_version from spacex where landing_outcome = 'Success (drone ship)' and payload_mass_kg_ > 400
```

```
 * ibm_db_sa://vry32322:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/bludb
Done.
```

Out[14]:

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

Total Number of Successful and Failure Mission Outcomes

- This query displays the total number of successful and failure mission outcomes

Task 7

List the total number of successful and failure mission outcomes

```
%sql select mission_outcome, count(*) from spacex group by mission_outcome
```

```
 * ibm_db_sa://vry32322:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:315
05/bludb
Done.
```

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- There were 12 boosters which have carried the maximum payload mass of 15600kg

```
8]: %sql Select Distinct booster_version, payload_mass_kg_ from spacex where payload_mass_kg_ = (select max(payload_m
```

```
 * ibm_db_sa://vry32322:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:315
05/bludb
Done.
```

8]:

booster_version	payload_mass_kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

2015 Launch Records

- The query displays 2 failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
]: %sql select landing_outcome, booster_version, launch_site from spacex where landing_outcome = 'Failure (drone shi
 * ibm_db_sa://vry32322:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:315
05/bludb
Done.
```

]:

landing_outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- This query ranks 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

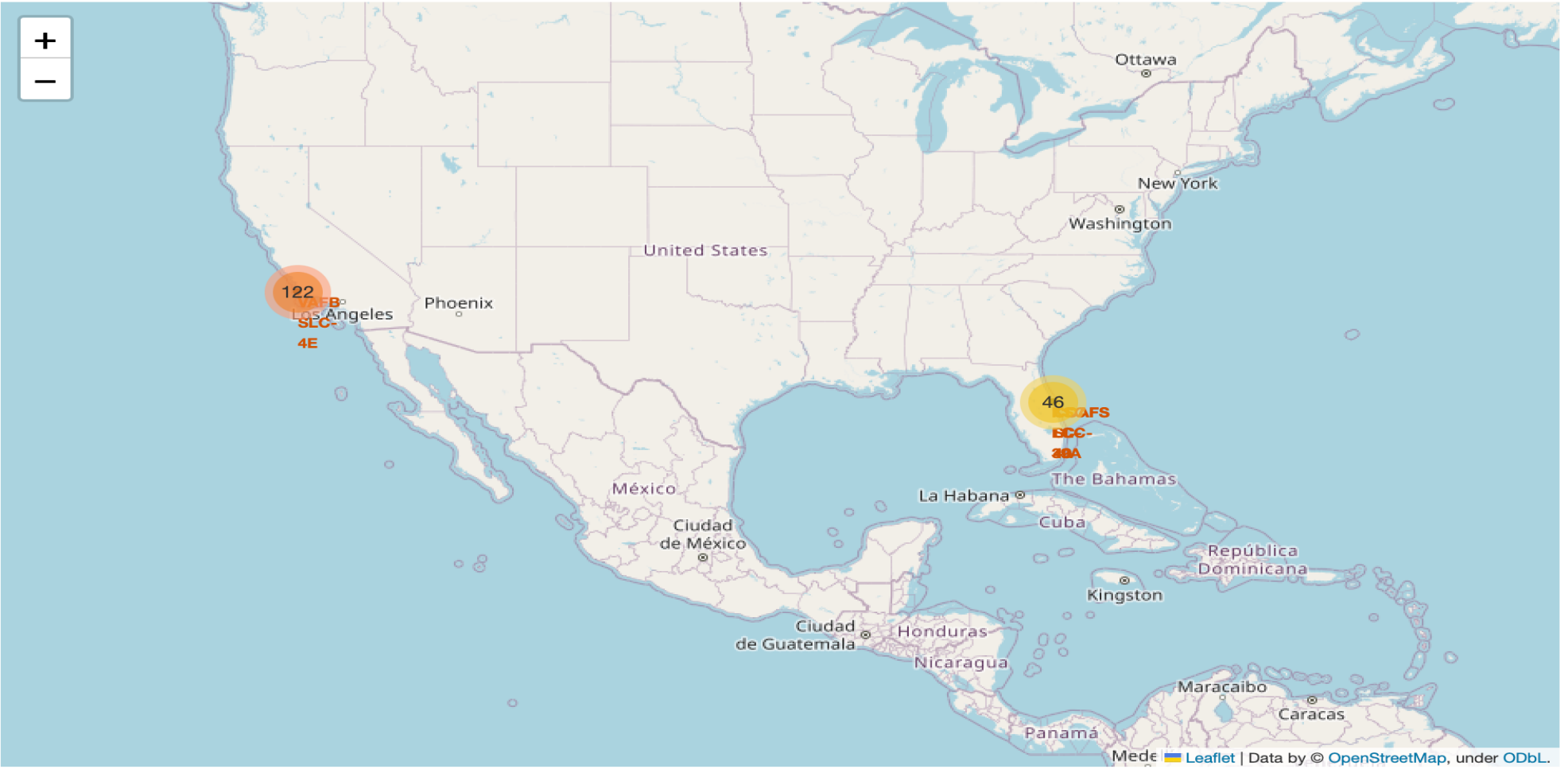
```
%sql select landing_outcome, count(*) as count from spacex where date >= '2010-06-04' and date <= '2017-03-20' gr
```

```
 * ibm_db_sa://vry32322:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:315
05/bludb
Done.
```

landing_outcome	COUNT
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


Section 3
Launch Sites
Proximities Analysis

All Launch Sites



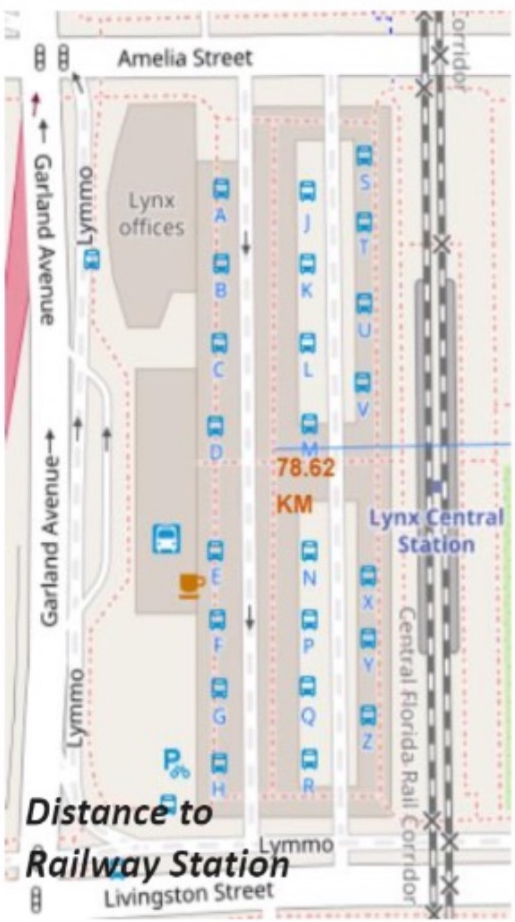
Launch sites are near sea, probably by safety, but not too far from roads and railroads

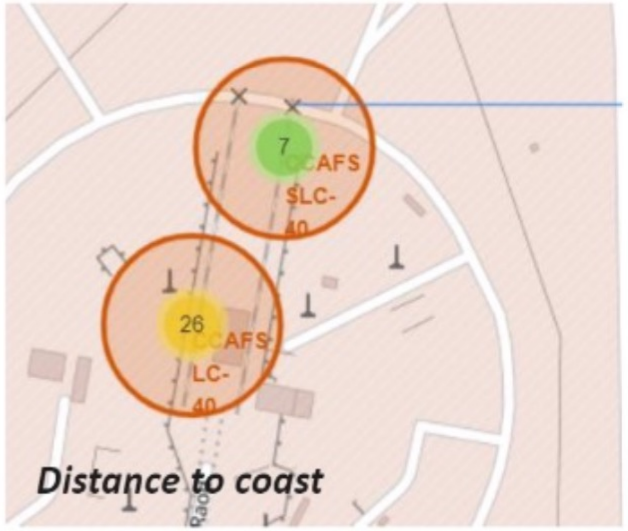
Launch Outcomes by Site

- Example of KSC LC-39A launch site launch outcomes

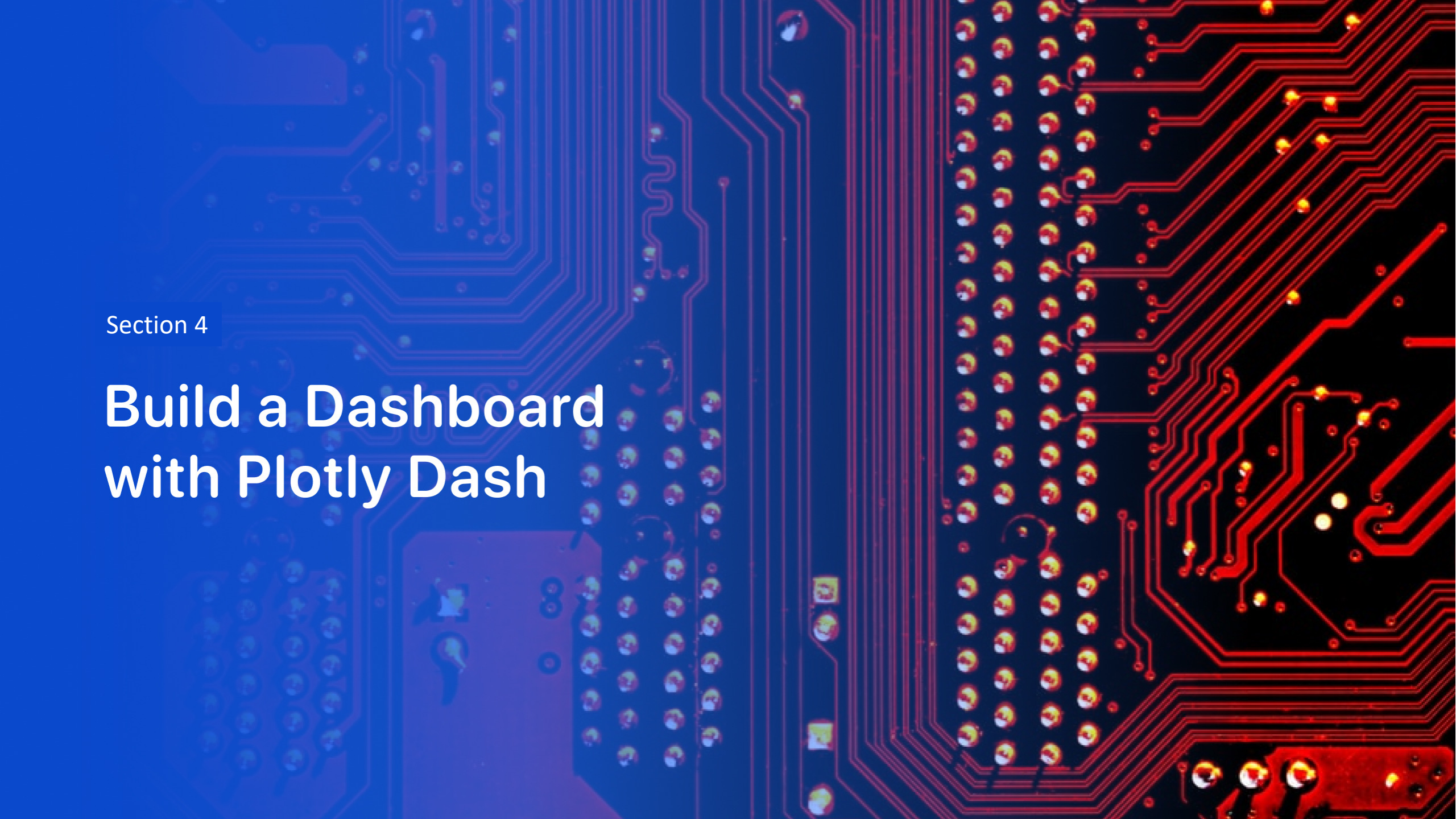
- Green markers indicate successful and red ones indicate failure

Launch Site distance to landmarks

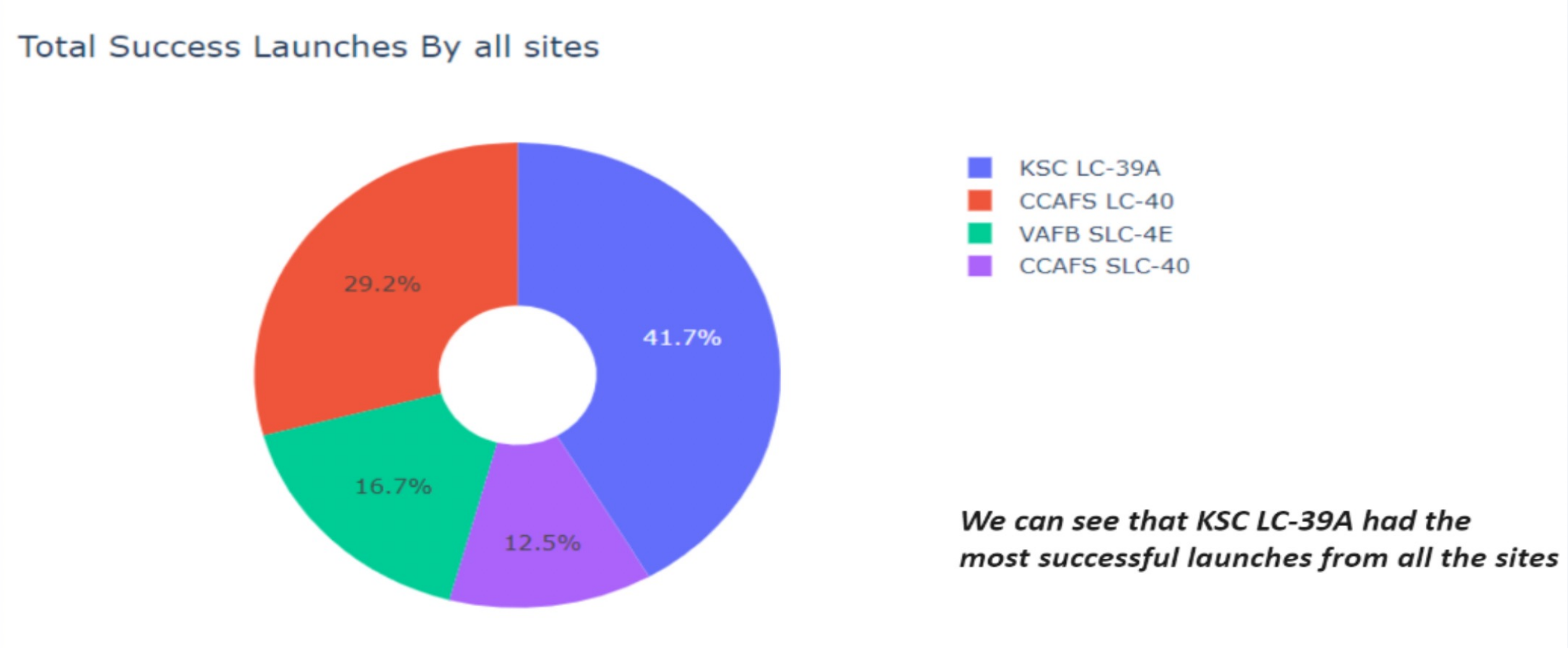




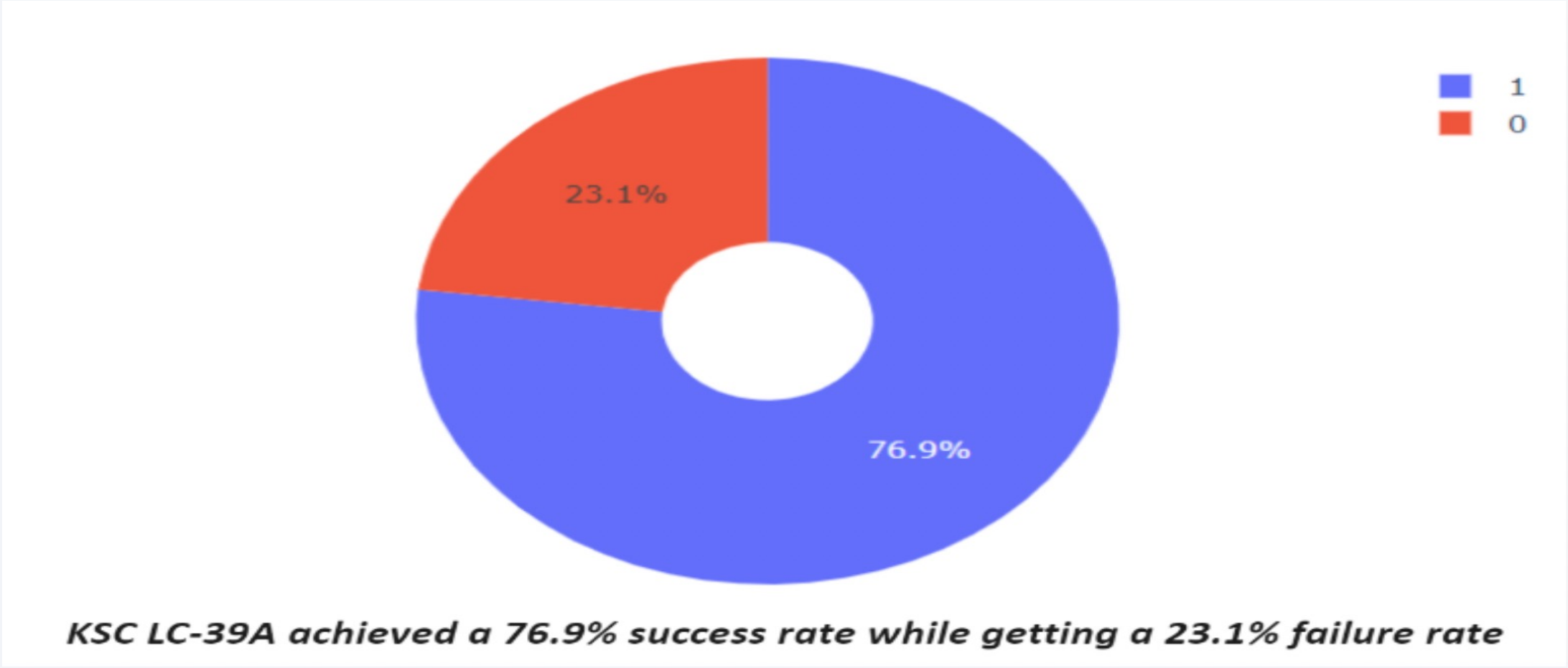
- Are launch sites in close proximity to railways? No
- Are launch sites in close proximity to highways? No
- Are launch sites in close proximity to coastline? Yes
- Do launch sites keep certain distance away from cities? Yes


Section 4
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site

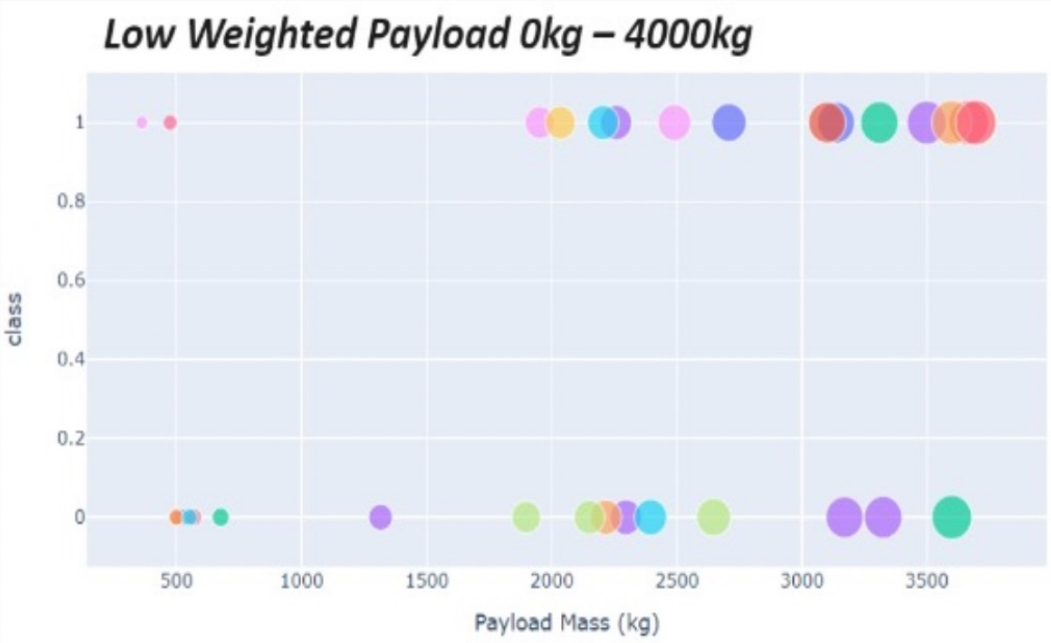


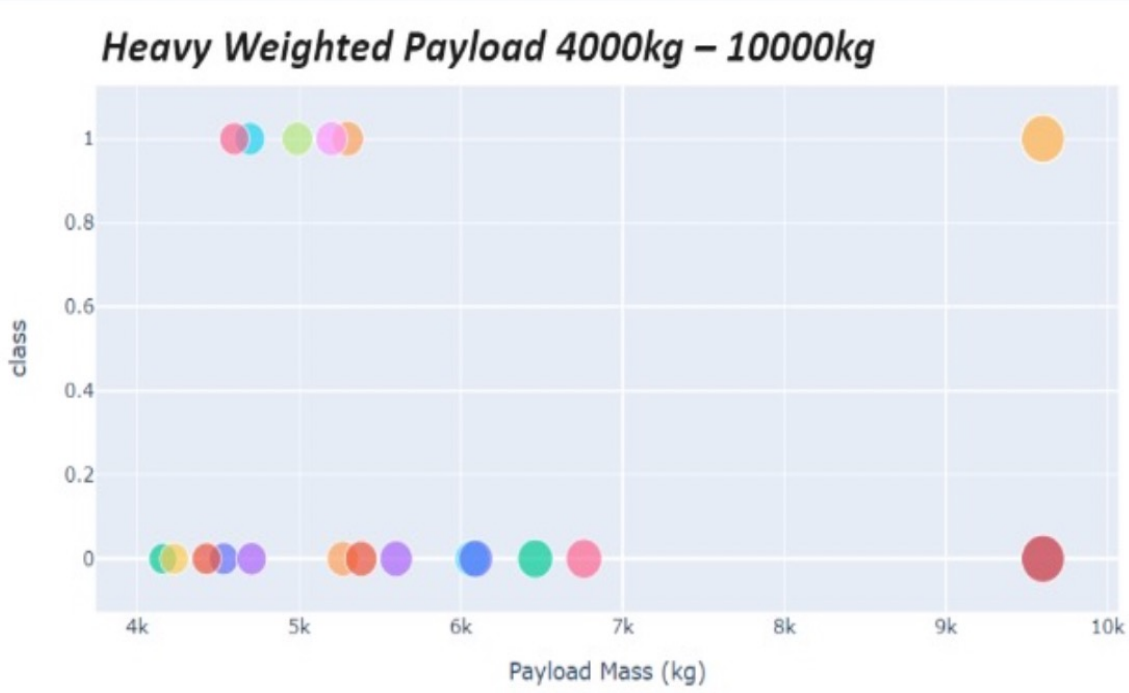
Pie chart showing the Launch site with the highest launch success ratio



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

Scatter plot of Payload vs Launch for all sites, with different payload selected in the range slider





We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

Classification Accuracy

The decision tree classifier is the model with the highest classification accuracy

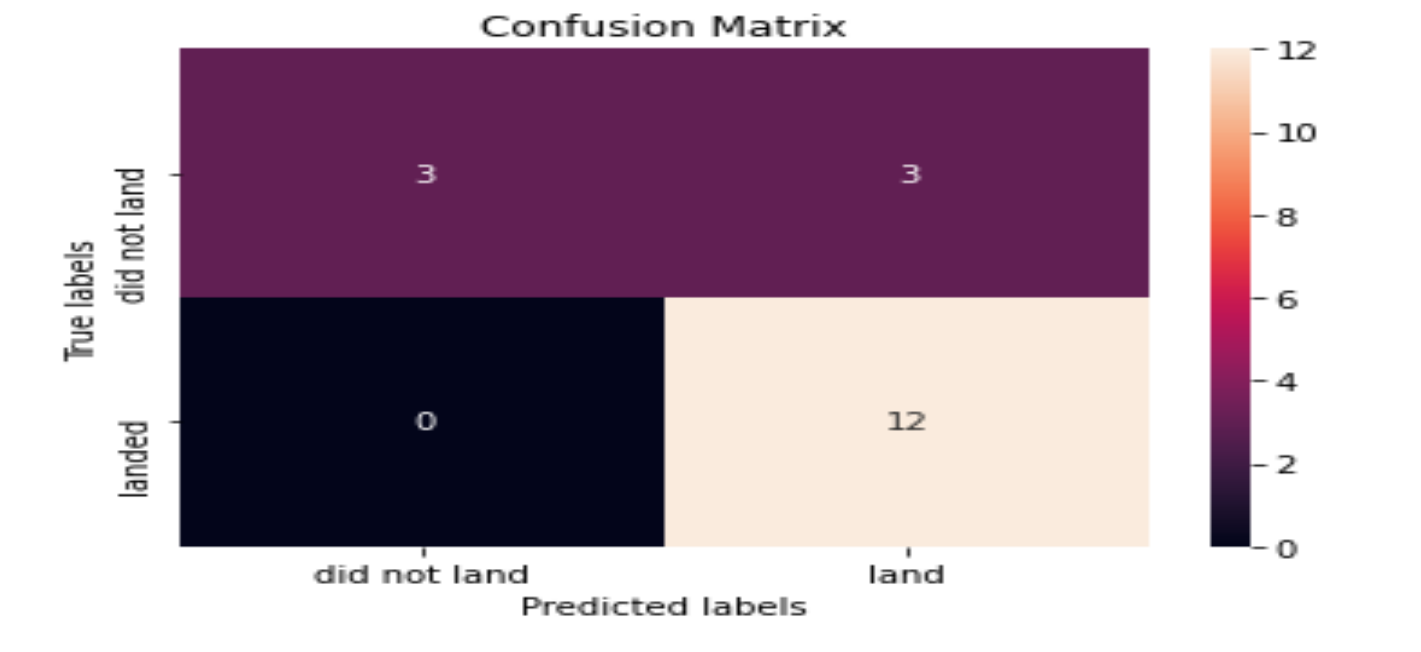
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8732142857142856
Best params is : {'criterion': 'gini', 'max_depth': 6, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'splitter': 'random'}
```

Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.
- ...

Thank you!