Winning Space Race with Data Science

IBM Data Science Capstone Project – Rocketing to Mars?

Wesley Burchnall
2022-01-05

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with Data Visualization
  - Exploratory Data Analysis with SQL
  - Building an Interactive Map with Folium API
  - Building a Plotly Dashboard
  - Predictive analysis (Classification)
- Summary of all results (presented in three ways)
  - Exploratory data analysis results
  - Interactive analytics demo via Screenshots
  - Predictive analysis result

Introduction

- Project background and context

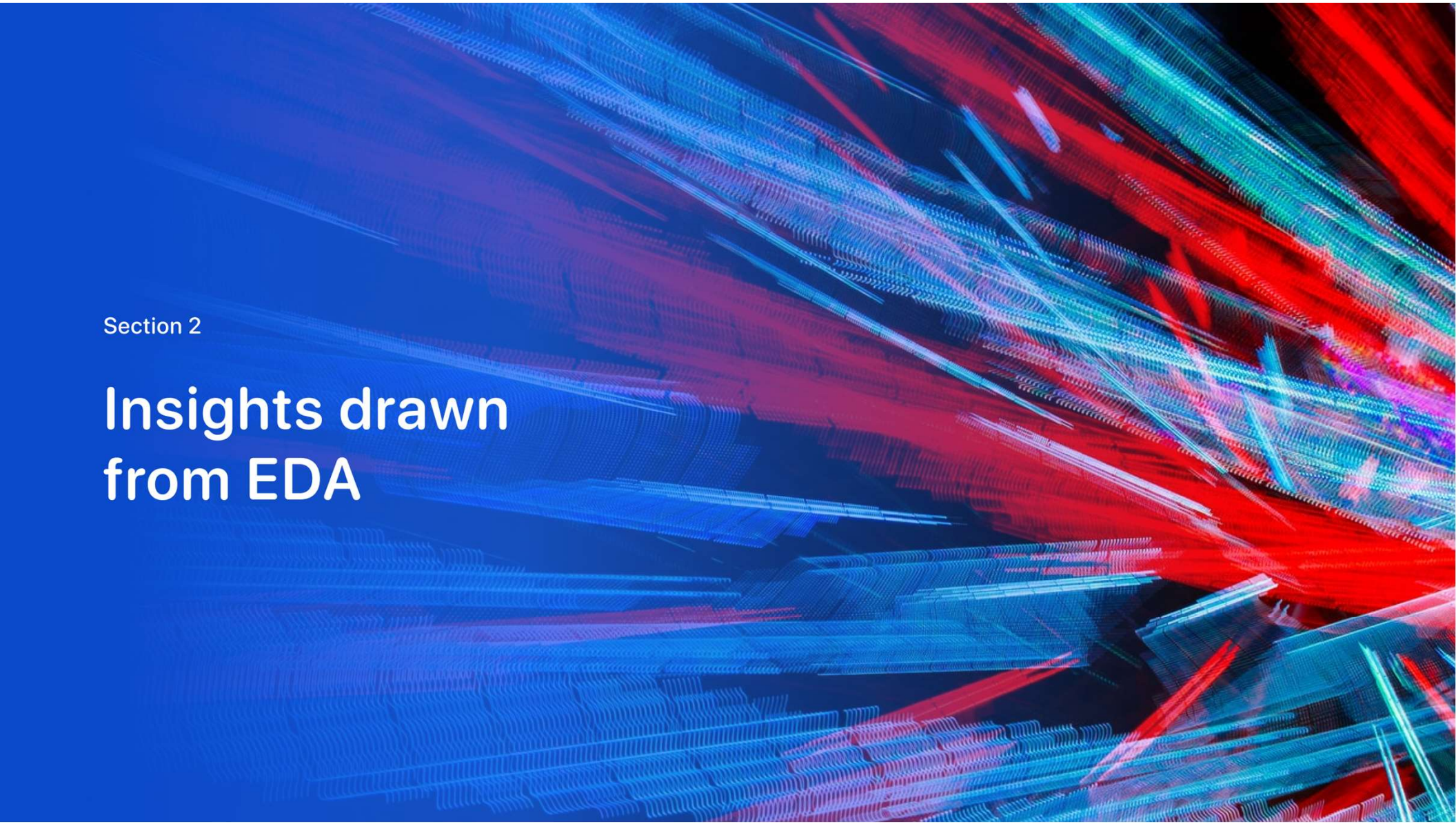
Results predicted included if a Falcon 9 first stage rocket will land successfully. SpaceX provides cost savings by having reusable first stage rockets which sets their price tag for a launch at $62 million USD versus other companies' $165 million USD. By the likelihood a first-stage rocket lands, we can determine appropriate pricing.

- Problems you want to find answers

What factors influenced when the rocket landed successfully?

What is the effect of these factors in determining the success rate of a successful landing?

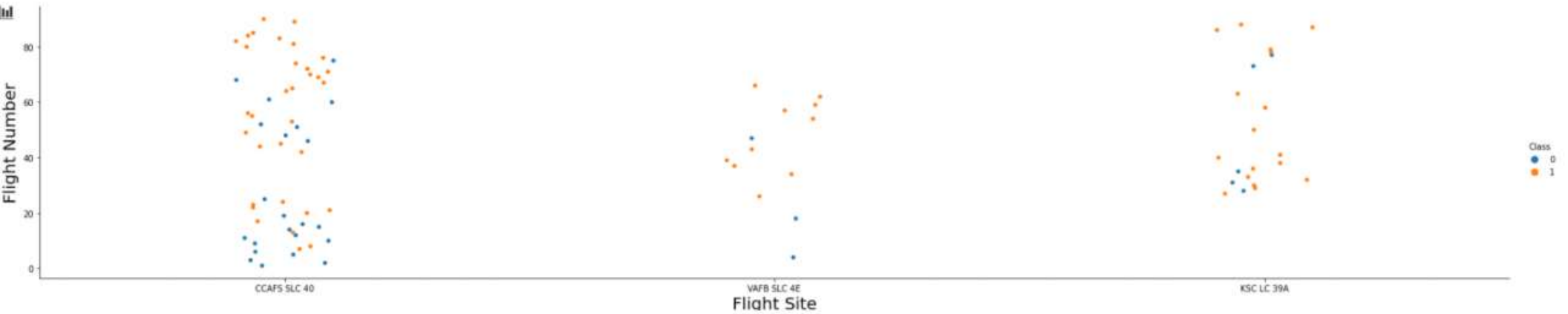
What conditions are necessary for SpaceX ensure the best rocket success landing rate for optimal profits?



Section 2

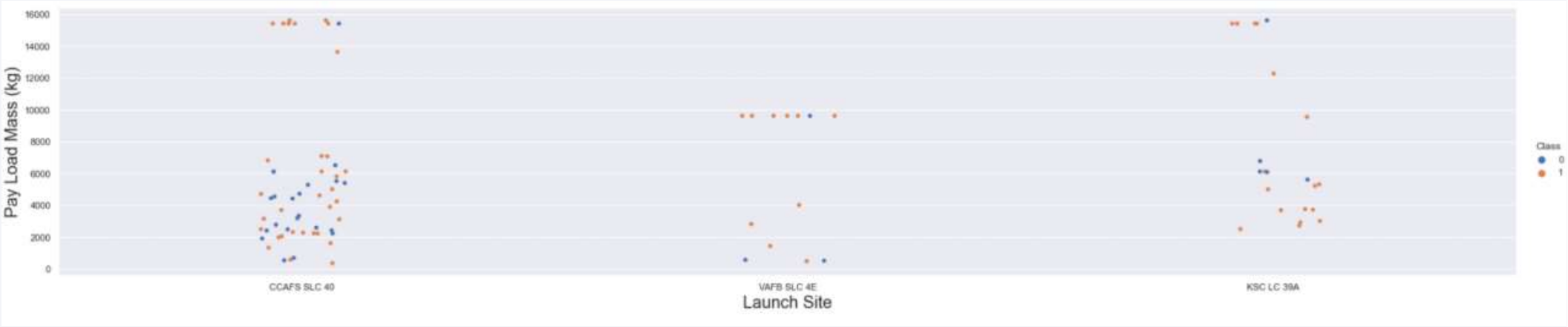
Insights drawn from EDA

Flight Number vs. Launch Site



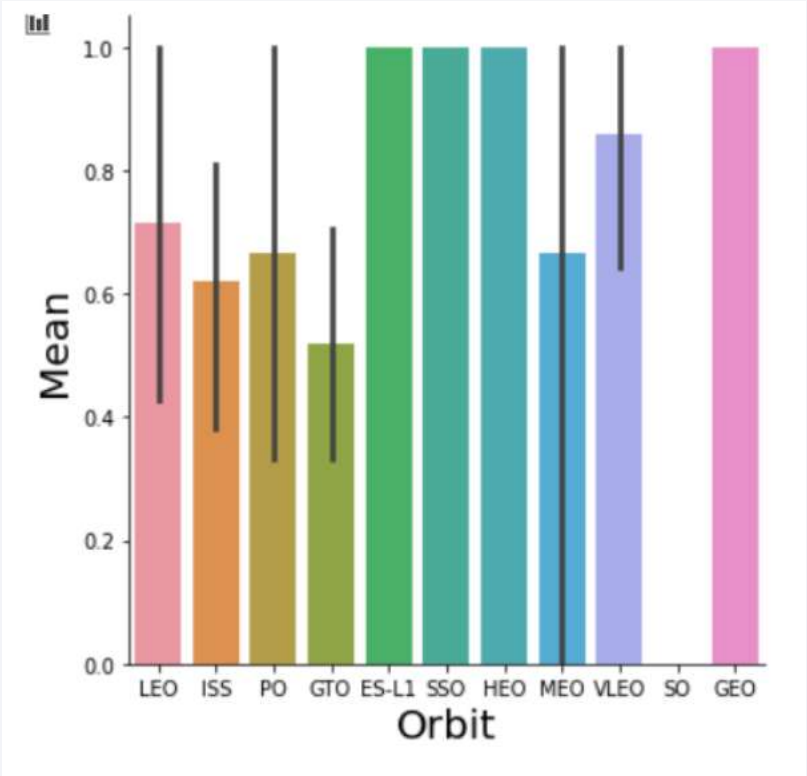
The higher the flight number at each test site, the great the likelihood the flight was successful. Most failures are clustered at the bottom/low flight numbers across all sites. At VAFB SLC 4E, the first two flights were failures and both flights had a low flight number. However, 10 out of 11 later flights were successful with higher flight numbers.

Payload vs. Launch Site



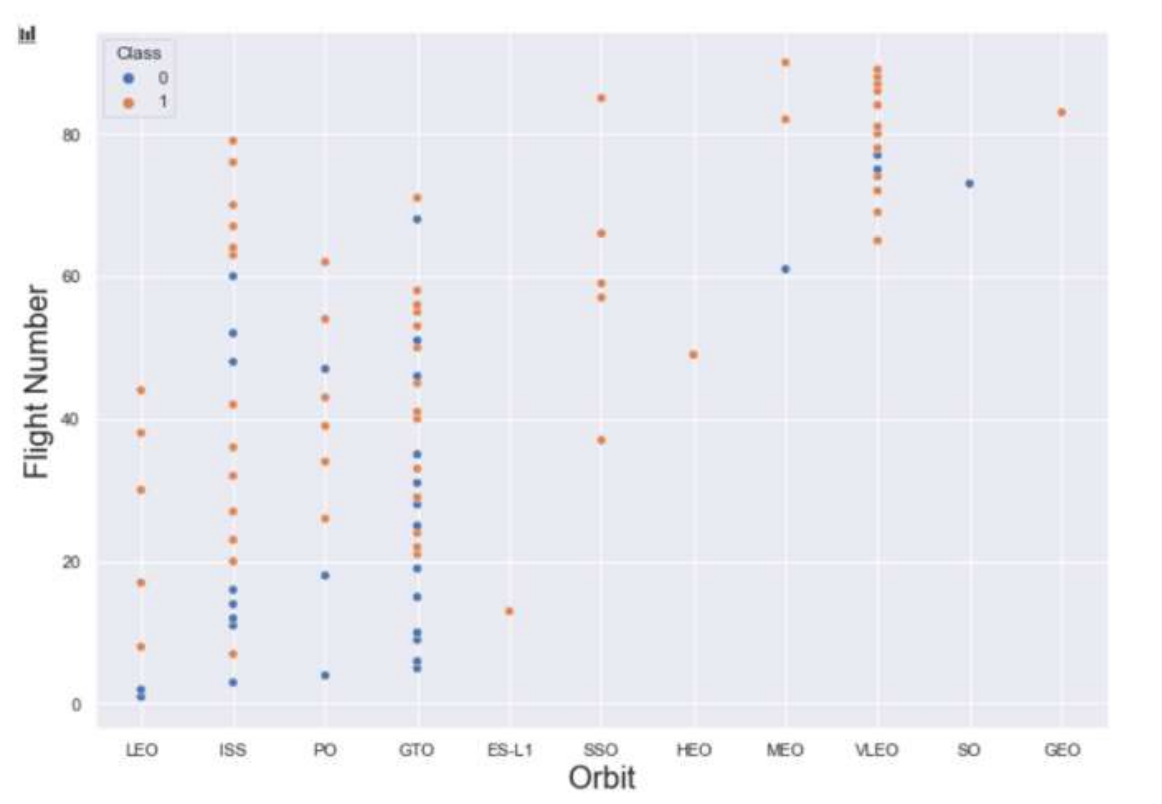
The correlation between payload and successful rocket launches changes based upon payload size. At smaller payloads, the ratio of successful launches to failures was low (~1:1 to ~2:1). These low-payloads might represent test-launches during development. During testing, one does not want to risk expensive cargo. At higher payloads, (likely after testing rocket designs), the success ratio of launches is much higher (~7:1 to ~8:1).

Success Rate vs. Orbit Type



The most successful orbitals were HEO, SSO, ES-L1 and GEO.
The least successful orbitals were ISS, PO and GTO.

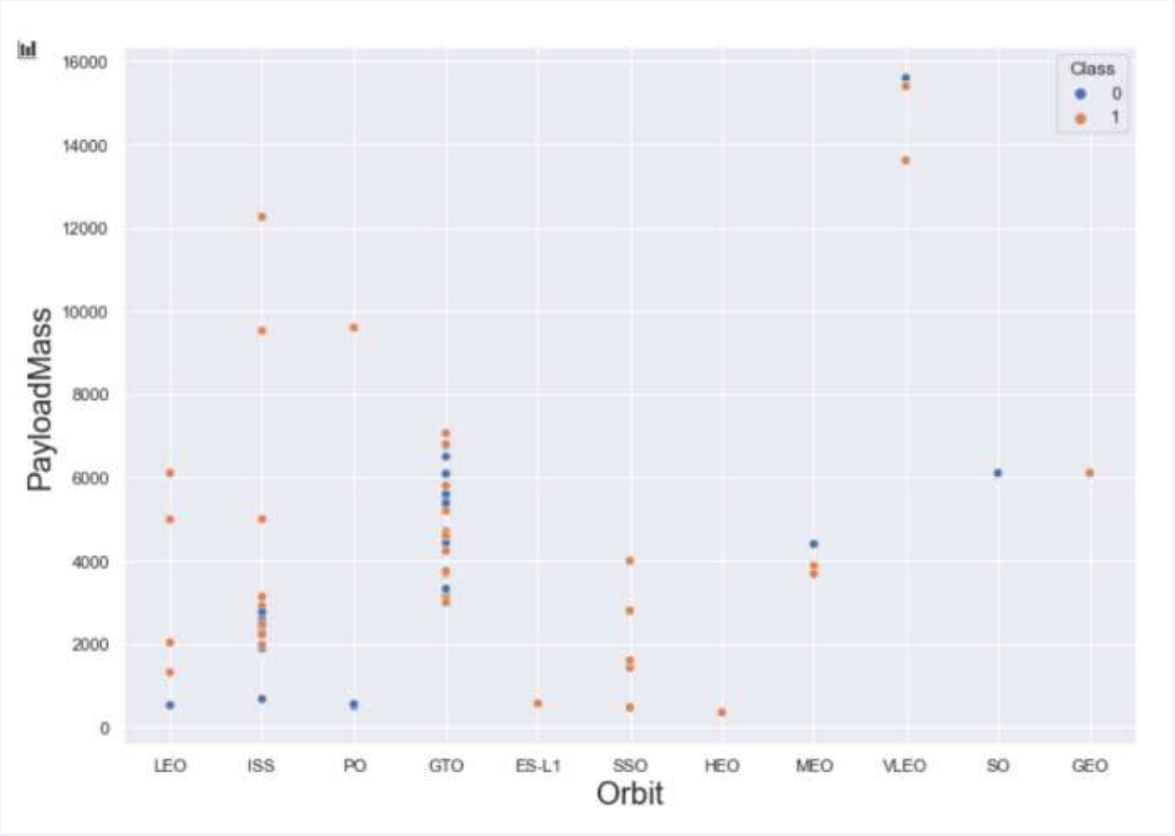
Flight Number vs. Orbit Type



The results of this plot are somewhat inconclusive. We can see for LEO that higher flight numbers were more successful. However, with ISS, PO, GPO and VLEO orbitals, it is harder to draw any clear conclusions or statistical inference between success and flight number.

We can observe that all ES-L1, SSO, HEO and GEO orbit launches were successful.

Payload vs. Orbit Type

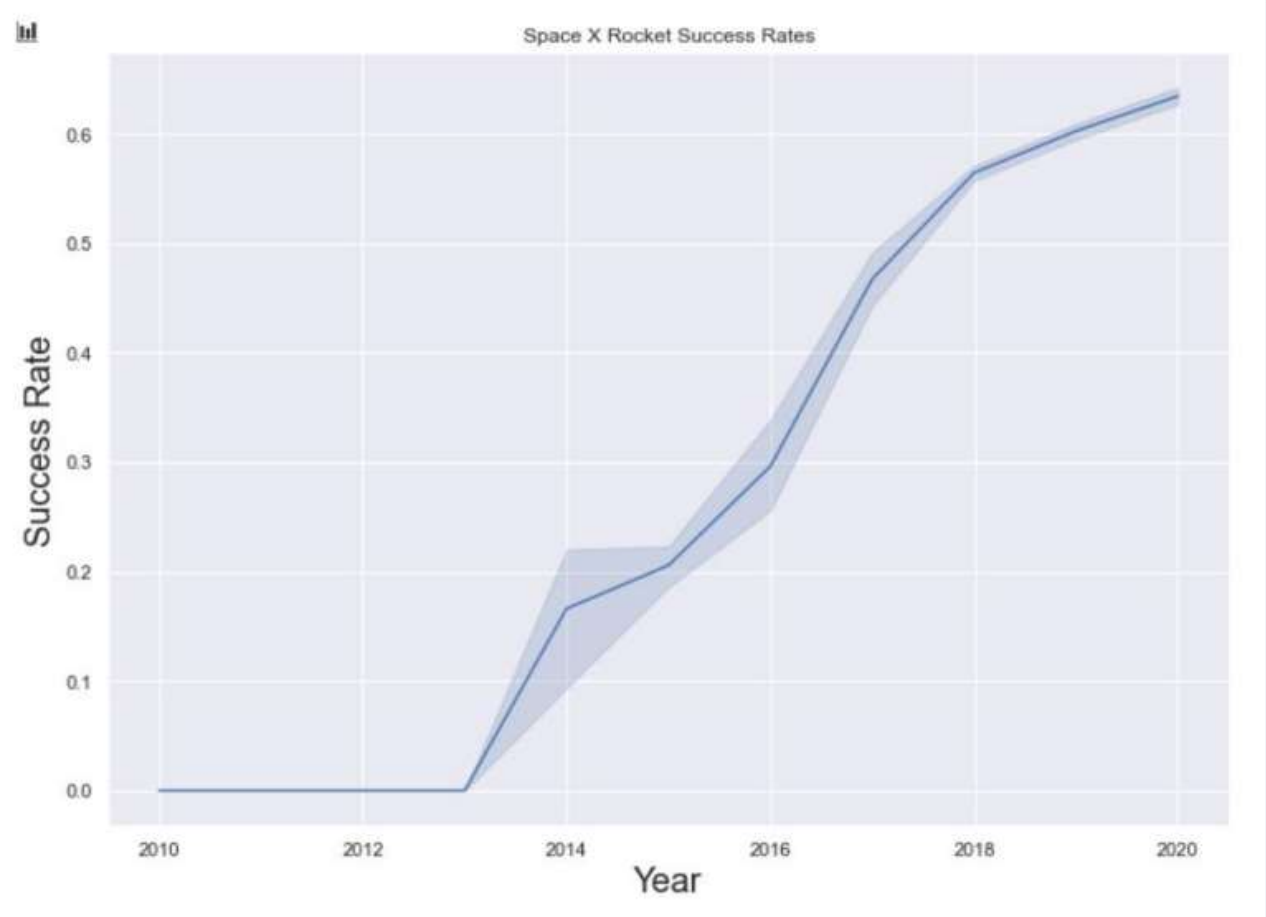


From this graph, we can see that at particular orbs like LEO and ISS, heavier payloads were associated with successful launches.

Where all ES-L1, SSO, HEO and GEO orbital launches were successful regardless of payload size.

The relationship between payload size and success in ISS, GTO, MEO and VLEO orbitals seems inconclusive. I suspect, the p value is quite high or r-squared value is quote low.

Launch Success Yearly Trend



As we would expect from other graphs, where higher flight numbers were associated with more successful launches, it appears over time, Space X has been improving at generating successful launches.

Only 18% of launches in 2014 were successful. By 2018, 56% were successful and by 2020, Space X had achieved a 63% success rate.

All Launch Site Names

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

Query used:

select DISTINCT Launch_Site from tblSpaceX

The word select indicates we are trying to retrieve records which are DISTINCT. DISTINCT gives us unique values/records from the column lunch_site from the table called tblSpaceX.

All Launch Site Names (Alternate Query)

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

Another query one can use to achieve the same output is:

%sql select Unique(LAUNCH_SITE) from SPACEXTBL;

The word select indicates we are trying to retrieve records which are “Unique” entries from the column gives in the brackets (LAUNCH_SITE) contained in the table SPACEXTBL.

Launch Site Names Begin with 'CCA'

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

Query used:

select DISTINCT LAUNCH_SITE from tblSpaceX where Launch_Site LIKE ‘CAA%’ LIMIT 5;

The word select indicates we are trying to retrieve records which are DISTINCT. DISTINCT gives us unique values/records from the column LAUNCH_SITE from the table called tblSpaceX. Adding where LAUNCH_SITE LIKE ‘CAA%’ implies the column LAUNCH_SITE must match a pattern beginning with CAA and be followed some string (%). The limit command reduces the results to only 5 records.

Total Payload Mass for NASA Launches

Total PayLoad Mass for Nasa: 45596kg

Query Used:

select SUM(PAYLOAD_MASS_KG_) TotalPayloadMass from tblSpaceX where Customer = 'NASA (CRS)'",'TotalPayloadMass

The SUM command adds up all the values in the column of the records retrieved. where Customer='NASA (CRS)' allows us to only select retrieve records when the customer was Nasa. Adding "TotalPayloadMass" is a column name that gets added to the output for readability.

Average Payload Mass by F9 v1.1

Average Payload Mass by F9 V1.1 rocket type is 2928.

Query Used:

select AVG(PAYLOAD_MASS_KG_) AveragePayloadMass from tblSpaceX where Booster_Version = 'F9 v1.1'

The avg command calculates the average of the column passed as a parameter from the table tblSpaceX and where the 'select'ed records returned are limited by 'where' the column Booster_Version is equal to the string 'F9 V1.1'

First Successful Ground Landing Date

First successful landing date of the drone ship was 22-12-2015.

Query Used:

select MIN(Date) SLO from tblSpaceX where Landing_Outcome = "Success (ground pad)"

The min locates the minimum value of the date-column passed as a parameter from the table tblSpaceX and where the 'select'ed records returned are limited by 'where' the column Landing_Outcome is equal to the string 'Success (ground pad)'

Successful Drone Ship Landing with Payload between 4000 and 6000

First successful landing date of the drone ship were:

F9 FT B1022
F9 FT B1026
F9 B4 B1021.2
F9 B4 B1031.2

Query Used:

select Booster_Version from tblSpaceX where Landing_Outcome = 'Success (drone ship)' AND Payload_MASS_KG_ > 4000 AND Payload_MASS_KG_ < 6000

The select command returns from table tblSpaceX 'select'ed records limited by 'where' the column Landing_Outcome is equal to the string 'Success (drone ship)' and the values in Payload_MASS_KG was between 4000 and 6000. It displays the Booster_Version that matches these conditions.

Total Number of Successful and Failure Mission Outcomes

- The query returned there were 100 Successful Missions and 1 Failed Mission.

Query used:

SELECT(SELECT Count(Mission_Outcome) from tblSpaceX where Mission_Outcome LIKE '%Success%') as Successful_Missions, (SELECT Count(Mission_Outcome) from tblSpaceX where Mission_Outcome LIKE '%Failure%') as Failure_Missions

This query uses the “LIKE” condition on a ‘where’ statement to search records were the column ‘Missing_Outcome’ contains the string “success” and returns the count of records as a custom-named column with the name Successful_Missions. The number of selected records is then counted with the “Count” command. The same approached is used to search and count for records where Mission_Outcome column contains the string “failure”.

Boosters Carried Maximum Payload

	Booster_Version	Maximum Payload Mass
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600

Query used:

```
SELECT DISTINCT Booster_Version, MAX(PAYLOAD_MASS_KG_) AS [Maximum Payload Mass] FROM tblSpaceX GROUP BY Booster_Version ORDER BY [Maximum Payload Mass] DESC
```

This query uses the DISTINCT to limit the returned records to unique-records of Booster_Version and uses the MAX command to return the maximum value found in the passed column. The Group By Booster_Version and ORDER BY server to create descending order of the results where the maximum payloads are shown at the top of the query. We can see the max payload was 15,600 Kg and where all F9 B5 type rockets.

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

- The total number of successful landing outcomes between 2010-06-04 and 2017-03-20 was 11 total (including “controlled” as success).

Query used:

select COUNT(LANDING_OUTCOME) as Count, LANDING_OUTCOME from tblSpaceX WHERE DATE BETWEEN '04-06-2010’ AND '20-03-2017’ GROUP BY LANDING_OUTCOME ORDER BY COUNT(LANDING_OUTCOME) DESC

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

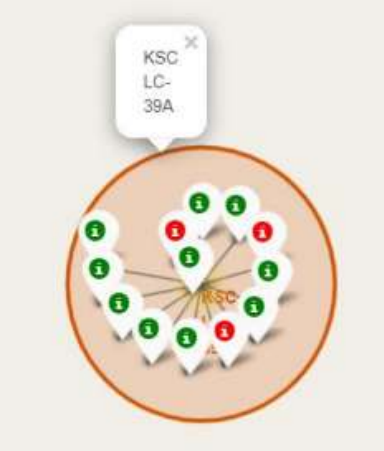
The COUNT(LANDING_OUTCOME) part of this query counts the quantity of landing outcomes and the records are limited to where the date is either greater than (after) ‘04-06-2010’ and less than (before) ’20-03-2017’ using the DATE BETWEEN command. Finally, GROUP BY LANDING_OUTCOME and BY COUNT DESC lists the outcomes by category and sorts in descending order

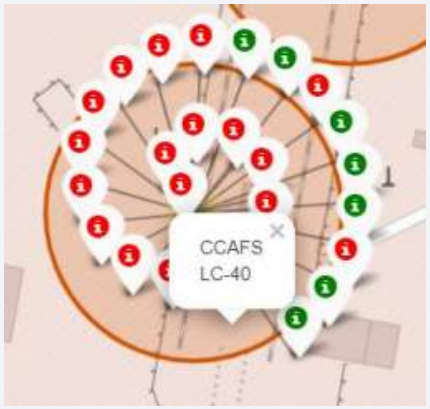

Section 4
Launch Sites
Proximities Analysis

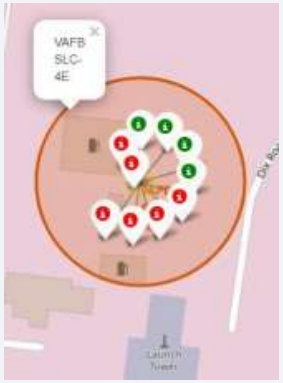
Global Placement of Launch Sites

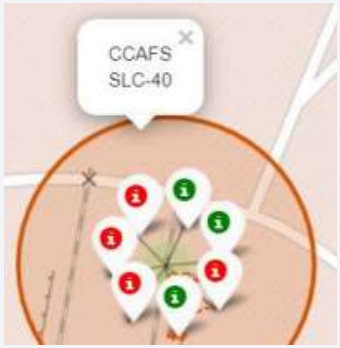
- From this images, we can see:
  - Several launch sites are in Florida on the East Coat of the United States
  - A launch site was in California on the West Coast of the United States
  - Elon Musk left California due to high tax rates and democratic polices and has moved his space-business almost entirely out of state.
  - No launch sites are located outside of the US boarder.

Clustered Labelled Markers by Color



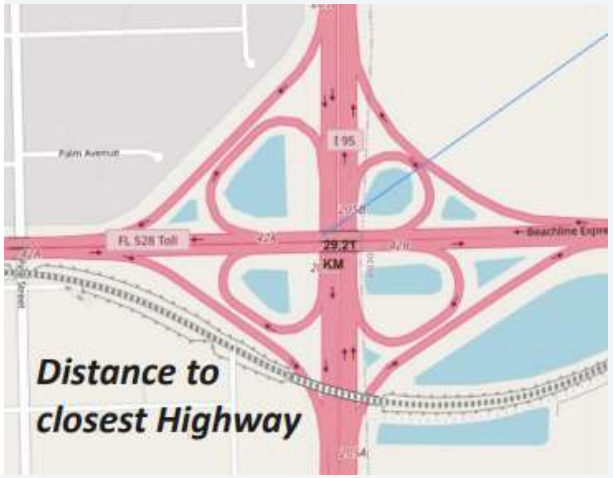


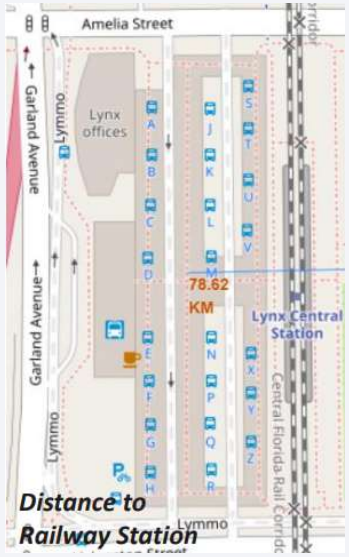




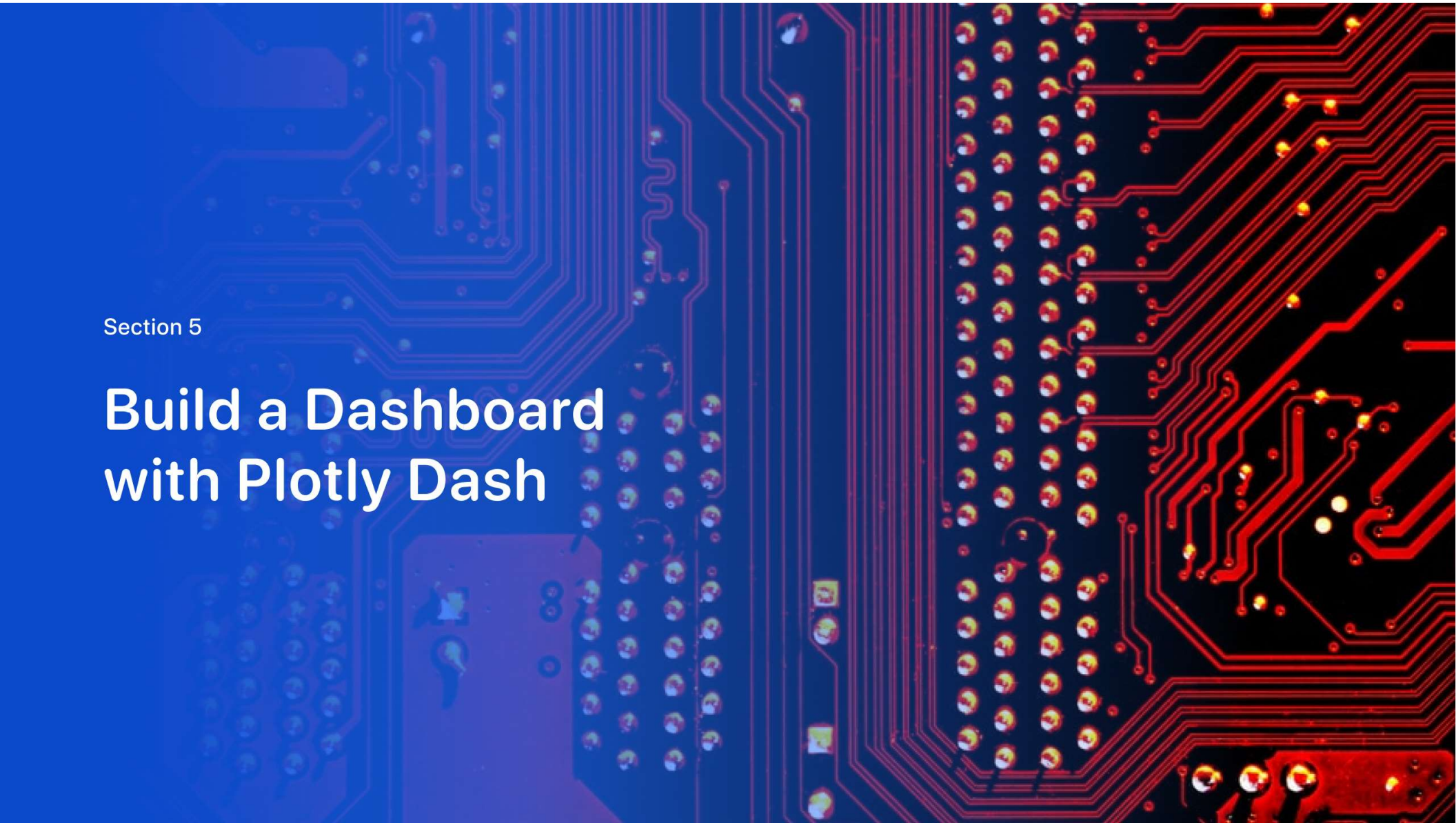
- From these images, we can see:
  - Successful Launches in Green
  - Failed Launches in Red
  - A Clockwise Spiral Pattern that orders the launches by flight number.
  - A small tag indicating the code of the launch site in white or orange. Ex. AFS LC-40 or CCAFS-LC 40

Launch Site Distances to Infrastructure





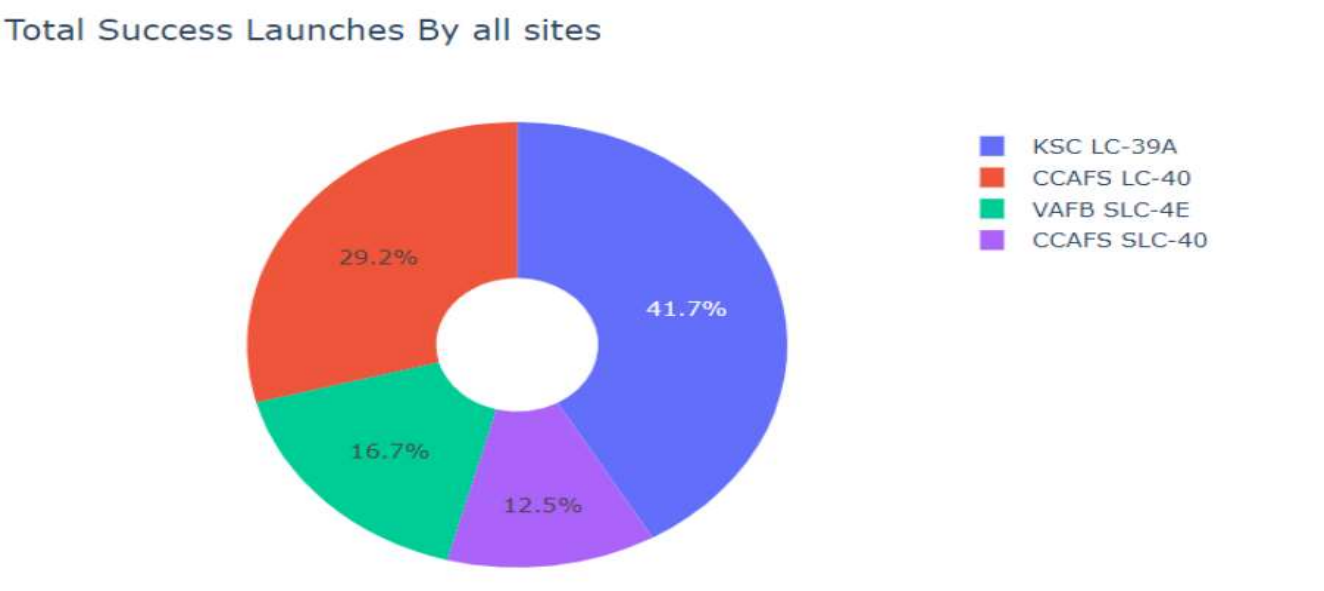
- From these images, we can see:
  - 21.9km to nearest highway
  - 0.90km to nearest coast
  - 78km to nearest city or railway station



Section 5

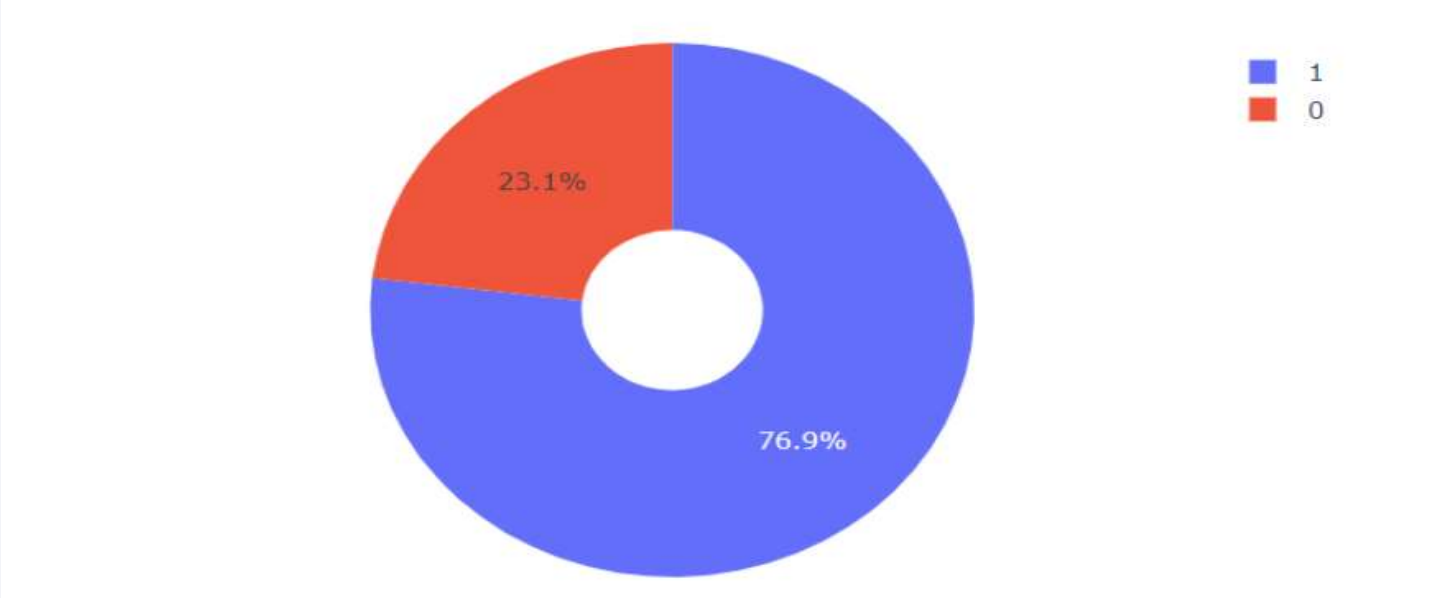
Build a Dashboard with Plotly Dash

Successful launches by Launch Site



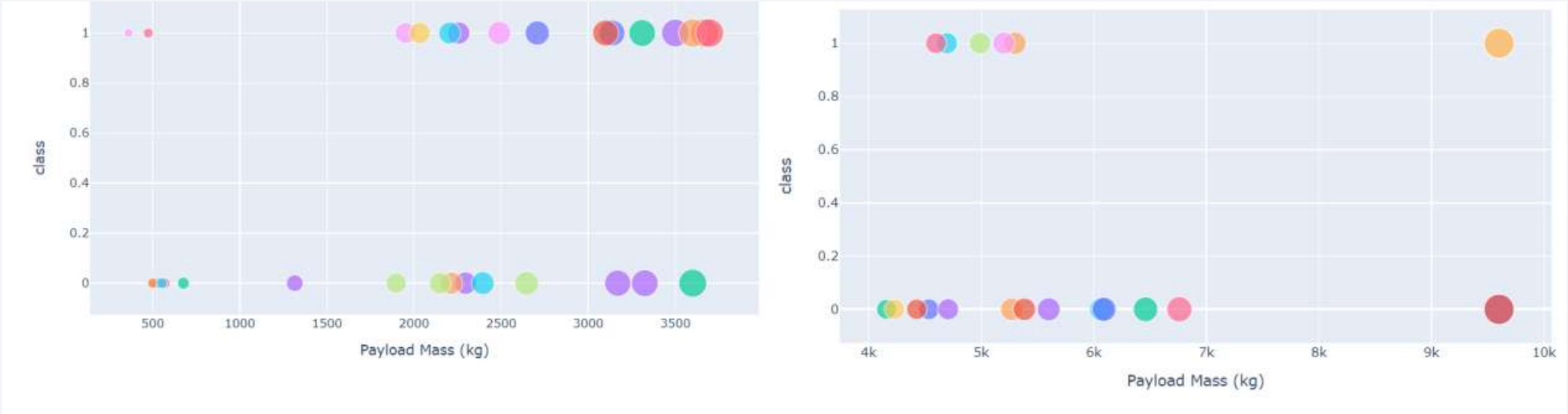
– KSC LC-39A wins as most successful launch site

Launch Statistics for the Most Successful Launch Site KSC LA-39A



KSC LC-39A had a 76.9% success ratio for launches.

Success outcomes by Payload



The leftmost graph shows launches with payloads under 4000kg and the rightmost shows payloads over 4000kg.
0 represents a failed launch. 1 indicates a successful launch. You can see a greater ratio of success to failure in payloads under 4000kg.

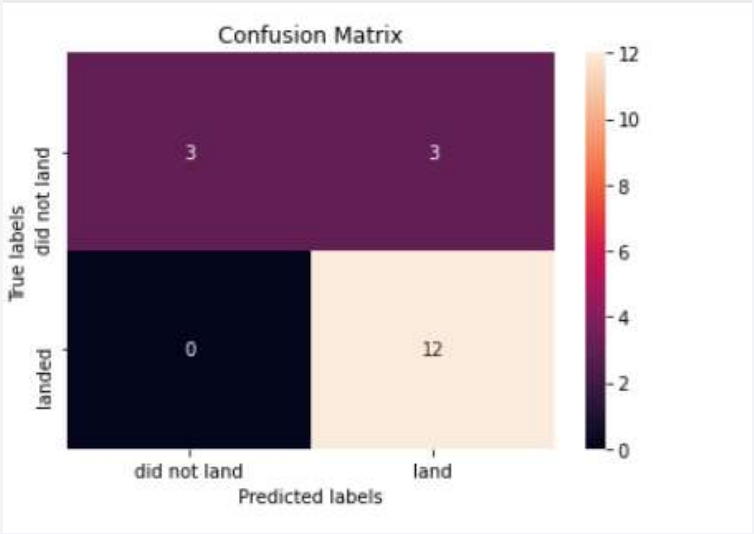
Section 6

Predictive Analysis (Classification)

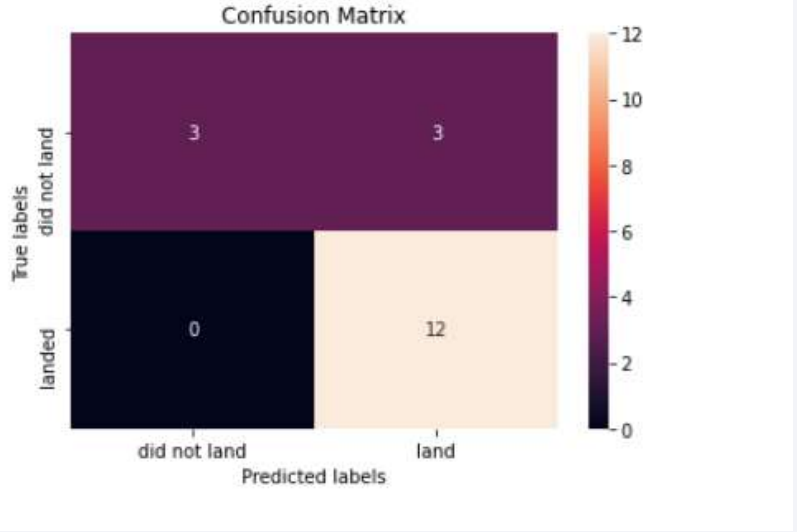
Confusion Matrix – Tied for Second Place

- Two models tied in performance with 84.82% training accuracy but these had poor test accuracy below 60%. These models are presented below. The predicted labels versus true labels shows the ratio of correct predictions compared to true-data. This indicates the model overfitted the training data and lost generalizability to predict new data.

```
yhat = knn_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```




```
yhat=svm_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Confusion Matrix – First Place

- The best test-performance was generated by Tree.

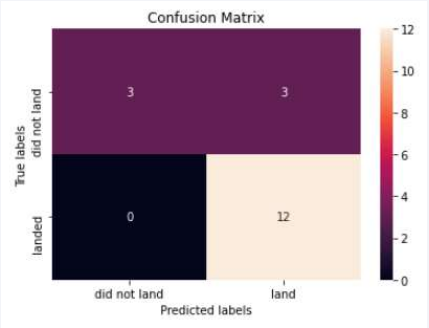
Calculate the accuracy of tree_cv on the test data using the method `score`:

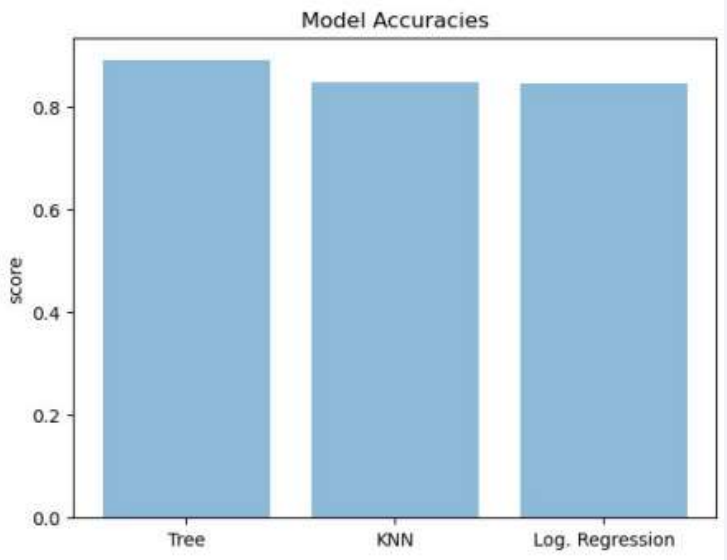
```
tree_cv.score(X_test, Y_test)
```

```
0.6666666666666666
```

We can plot the confusion matrix

```
yhat = tree_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```






```
print("tuned hpyerparameters :(best parameters) ",tree_cv.best_params_)
print("accuracy :",tree_cv.best_score_)
```

```
tuned hpyerparameters :(best parameters)  {'criterion': 'entropy', 'max_depth': 8, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 1
0, 'splitter': 'random'}
accuracy : 0.8857142857142856
```

Conclusions

- From the classification model results, tree_cv model had the best classification training accuracy and score during my attempts to train and test the data.
- From the interactive visual analysis with the plotly dashboard, we can see lower payloads increased the chances of success compared to higher payloads.
- From EDA with data visualization, successful rate of launch improved over time in years indicating they may eventually near-perfect launch rates. 2018 was a setback year and did not perform as well as neighboring years.
- From interactive visual analysis with plotly dashboard, KSC LC-39A had the more successful launches out of all the launch sites.
- From EDA with SQL results, the best success rates in terms of orbits was GEO, HEO, SSO and ES-L1 in that order.

Thank you!