IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hei Yin Kyle Chan
2/10/2022

Outline

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

Executive Summary

- I built a machine learning model to predict the success/failure of Falcon 9 landings
- EDA and subsequent analyses show that orbit type and payload mass are key factors in determining mission outcome
- With a simple decision tree classifier, the model was able to predict 83.33% of the out-of-sample cases correctly.

Introduction

- According to SpaceX, Falcon 9 cost around $62 mil, which is less than half of other rocket providers
- The cost saving comes from the fact that the first-stage Falcon 9 rockets can be reused
- Business Problem Statement:
  - *How costly is a Falcon-9 rocket launch?* If we can predict whether the first stage will land, we can determine the cost of a launch easily.
  - *How can we improve the success rate of the first-stage?* If we can improve further, we might be able save even more by increasing the re-use rate of the first-stage rockets.

Section 1

Methodology

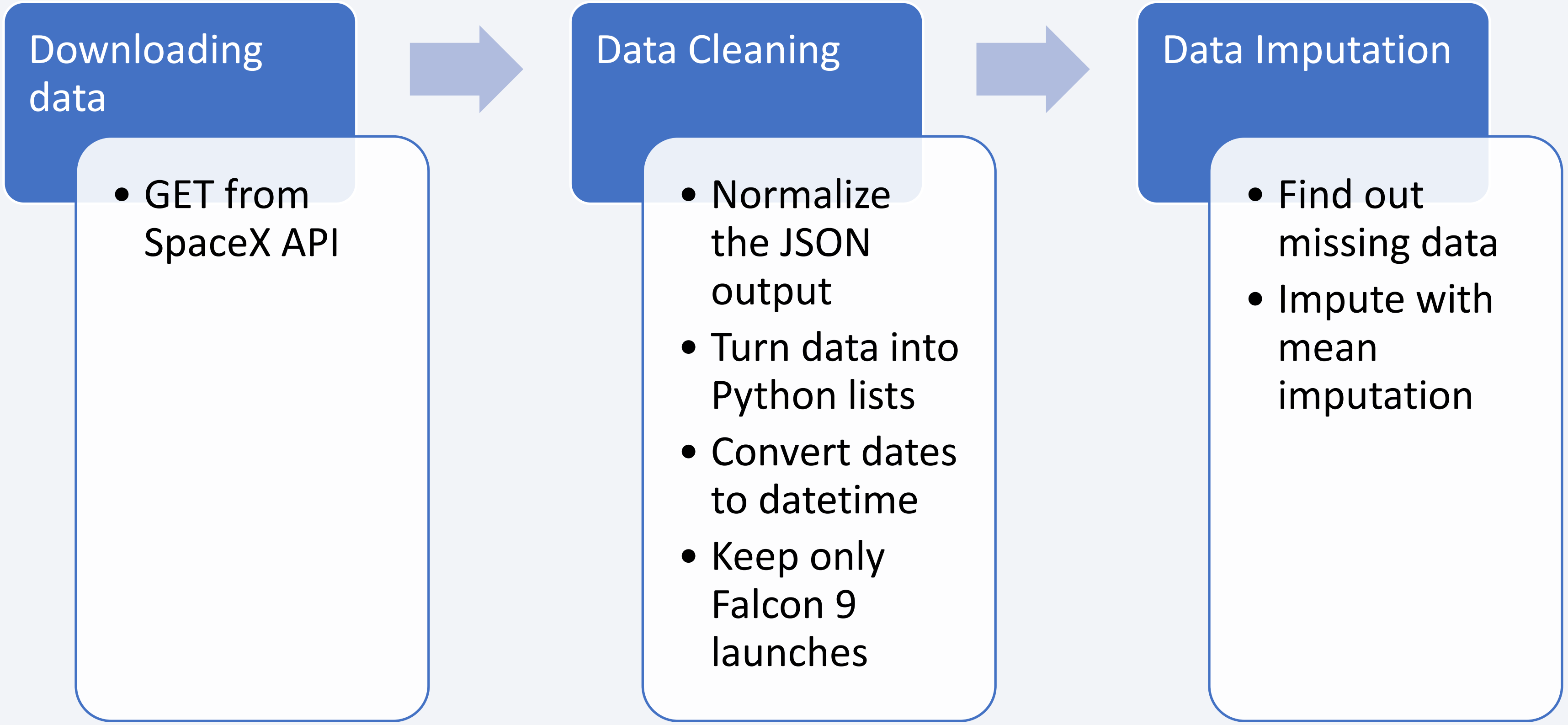
Methodology

Executive Summary

- Data Collection: I gathered the data from the SpaceX official API. The API returned 91 instances of Falcon 9 First-stage tests from 2010 to 2020.
- Data wrangling: I replaced missing values using the mean imputation method.
- Exploratory data analysis (EDA): I used SQL queries to conduct the EDA, and learnt that payload mass is an important factor.
- Interactive Visual Analytics: I constructed a Folium map to visualize the launch sites, and an interactive dashboard (Plotly) showing launch statistics.
- Classification:
  - Four models were chosen to fit the data
  - Models are tuned by Grid-search with Cross Validation
  - Eventually, the models perform more or less the same. I used a decision tree as the final model.

Data Collection

- The Falcon 9 data is collected from the SpaceX API. (https://api.spacexdata.com/). The data collection and wrangling process is illustrated below:



Downloading data

- GET from SpaceX API

Data Cleaning

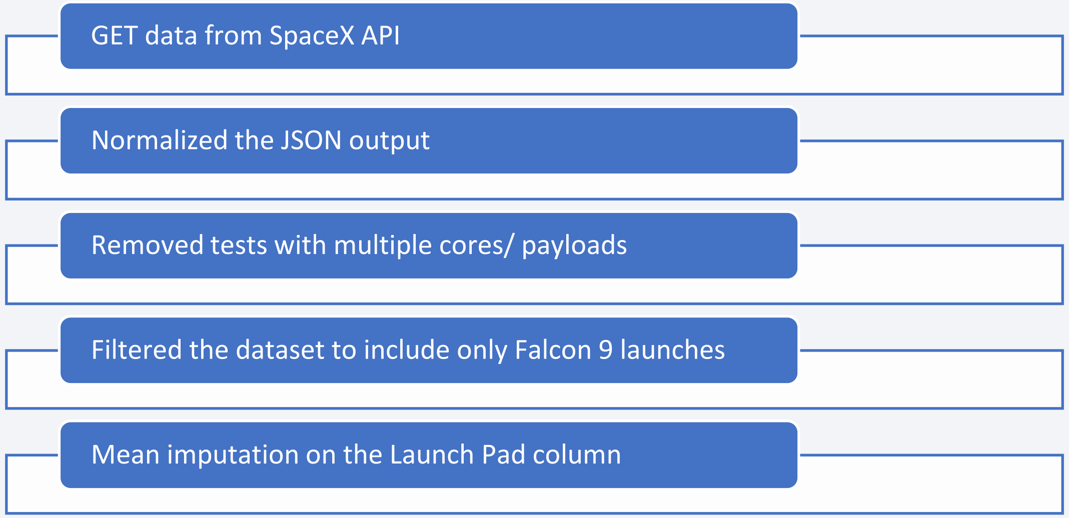
- Normalize the JSON output
- Turn data into Python lists
- Convert dates to datetime
- Keep only Falcon 9 launches

Data Imputation

- Find out missing data
- Impute with mean imputation

Data Collection – SpaceX API

- We used the GET command from the SpaceX API
- Conducted some data wrangling such as data imputation and filtering of data.
- This results in 90 launches in the dataset.
- Jupyter Notebook Reference Link: https://github.com/kylechanpols/IBM_Datascience_Cert/blob/main/01-jupyter-labs-spacex-data-collection-api.ipynb



GET data from SpaceX API

Normalized the JSON output

Removed tests with multiple cores/ payloads

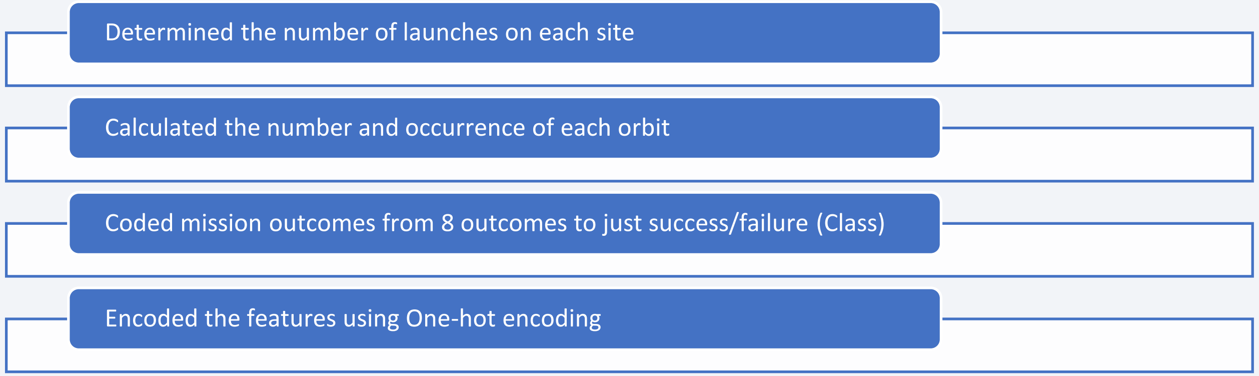
Filtered the dataset to include only Falcon 9 launches

Mean imputation on the Launch Pad column

Data Wrangling

- Mission outcomes are further coded, breaking down into simpler categories

- Reference: https://github.com/kylechanpols/IBM_Datascience_Cert/blob/main/03-labs-jupyter-spacex-Data%20wrangling.ipynb



Determined the number of launches on each site

Calculated the number and occurrence of each orbit

Coded mission outcomes from 8 outcomes to just success/failure (Class)

Encoded the features using One-hot encoding

EDA with Data Visualization

- An EDA is conducted by visualizing the correlation between outcomes and features.
- More details can be found in the following list

- Check Falcon-9 Progress: Plot Flight Number (history) against payload mass/ success.
- Check Launch site vs. mission outcomes to see if there are geographical trends determining success
- Check orbit type vs mission outcomes to see if orbit types determine success
- Check trends between flight number & orbit type to see if orbit strategy has changed over time
- Ditto, check if switching orbit strategy would lead to better mission outcomes

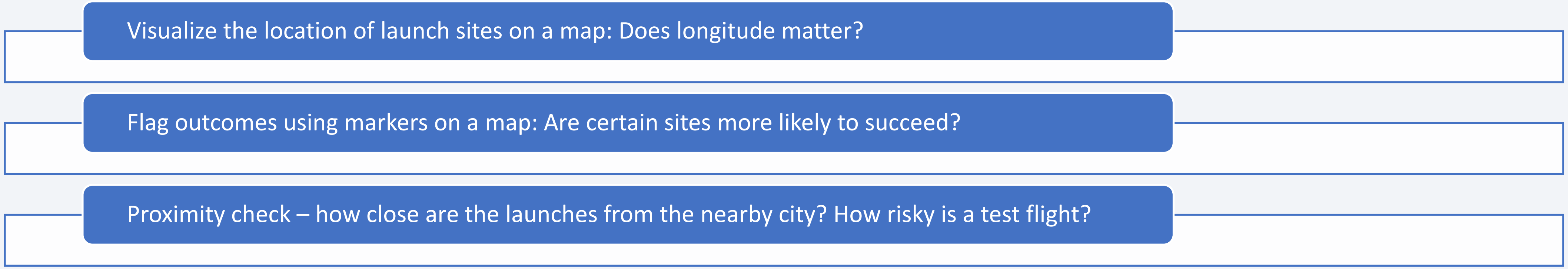
Reference: https://github.com/kylechanpols/IBM_Datascience_Cert/blob/main/03-labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with SQL

- In SQL, I acquired the following summaries:
  - `Display the names of the launch sites`
  - `Display the 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 of the first successful landing.`
  - `List the successful drone ships with a average payload mass (4000-6000 kg)`
  - `List the total number of successful and failure mission outcomes`
  - `List the names of the booster_versions which have carried the largest payload.`
  - `List the failed landing_outcomes in 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 June 2010 and March 2017.`
- Reference: https://github.com/kylechanpols/IBM_Datascience_Cert/blob/main/04-jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- The following tasks were done on an interactive map with Folium to check if there are geographical factors determining mission outcomes.



Visualize the location of launch sites on a map: Does longitude matter?

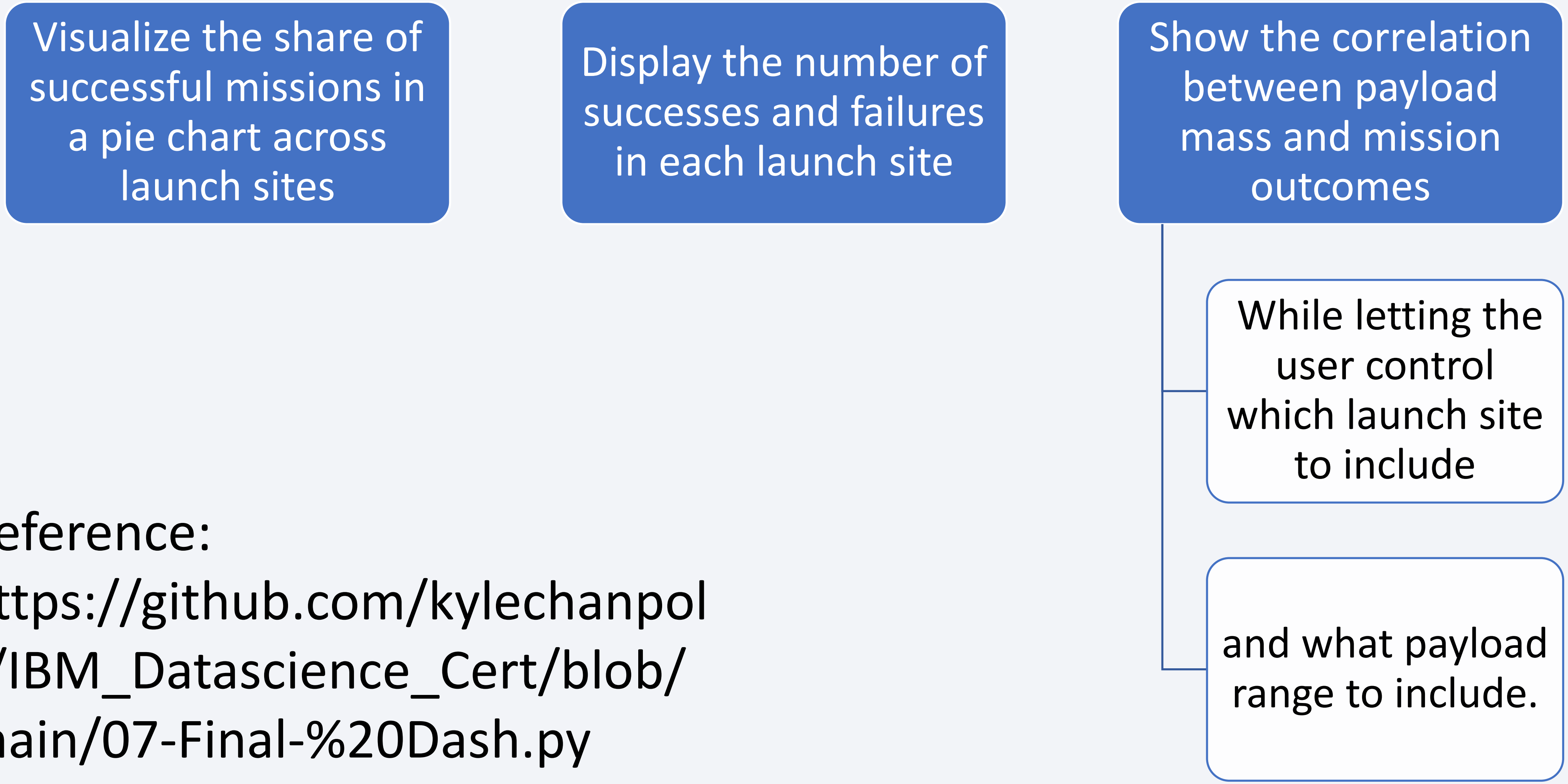
Flag outcomes using markers on a map: Are certain sites more likely to succeed?

Proximity check – how close are the launches from the nearby city? How risky is a test flight?

- Reference (Folium maps are disabled on Github – use the IBM Cloud Link to enable the Folium Maps) https://github.com/kylechanpols/IBM_Datascience_Cert/blob/main/06-Final%20Capstone%20-%20Dashboard.ipynb
- https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/1b101bbb-59ab-4e75-bc79-a969facda46d/view?access_token=69b0ea8214f5334ee0c67469782c4decdfbbe57d5aa791eeecc9616a78cdeb35

Build a Dashboard with Plotly Dash

- The following tasks were done on an interactive dashboard with Plotly/Dash.



Visualize the share of successful missions in a pie chart across launch sites

Display the number of successes and failures in each launch site

Show the correlation between payload mass and mission outcomes

While letting the user control which launch site to include

and what payload range to include.

- Reference: https://github.com/kylechanpols/IBM_Datascience_Cert/blob/main/07-Final-%20Dash.py

Predictive Analysis (Classification)

The model is developed using the following process:

- Standardized all data and turned them into numpy arrays
- Conducted Train/Test split with a 8:2 ratio
- Using Grid Search CV on a Logistic Regression Classifier and refit using best parameter found
- Using Grid Search CV on a Support Vector Machine Classifier and refit using best parameter found
- Using Grid Search CV on a Decision Tree Classifier and refit using best parameter found
- Using Grid Search CV on a K Nearest Neighbor Classifier and refit using best parameter found
- Found all models to perform somewhat similarly; Chose the Decision Tree Classifier for simplicity

Results (Executive Summary)

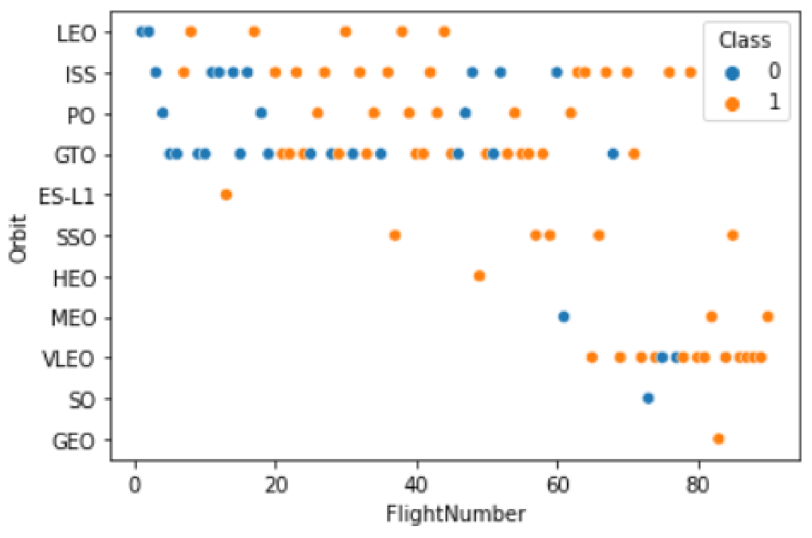
- EDA shows that orbit type and payload mass are key factors in determining mission outcome
- The Logistic Regression Model, Support Vector Machine and K-Nearest Neighbor models all predicted 83.33% of the outcomes correctly, with some false positives predicted
- The Decision Tree model was only able to predict around 66.67% of the outcomes correctly.

Section 2

Insights drawn from EDA

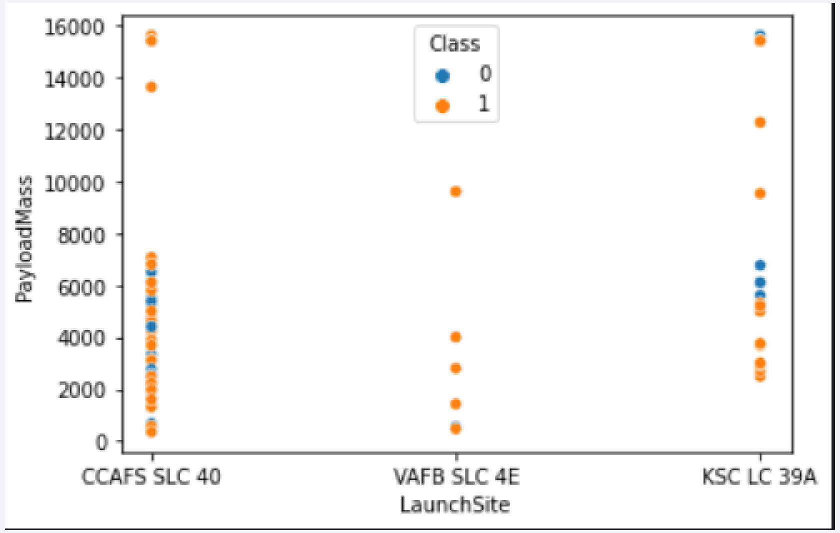
Flight Number vs. Launch Site

- Improvement in launch success – over time (I.e. larger flight number) there are more successes than failure.



Payload vs. Launch Site

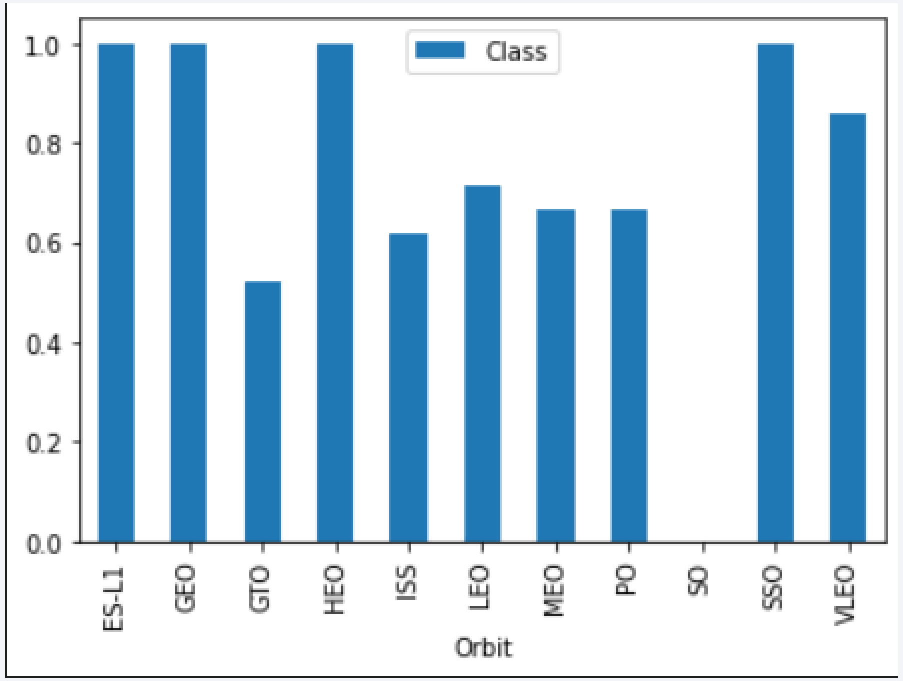
- CCAFS-SLC 40 carried out tests in all ranges, whereas KSC LC-39A's tests are clustered in the mid-range (4000-6000).
- VAFB SLC-4E carried out tests with lighter payloads.



Success Rate vs. Orbit Type

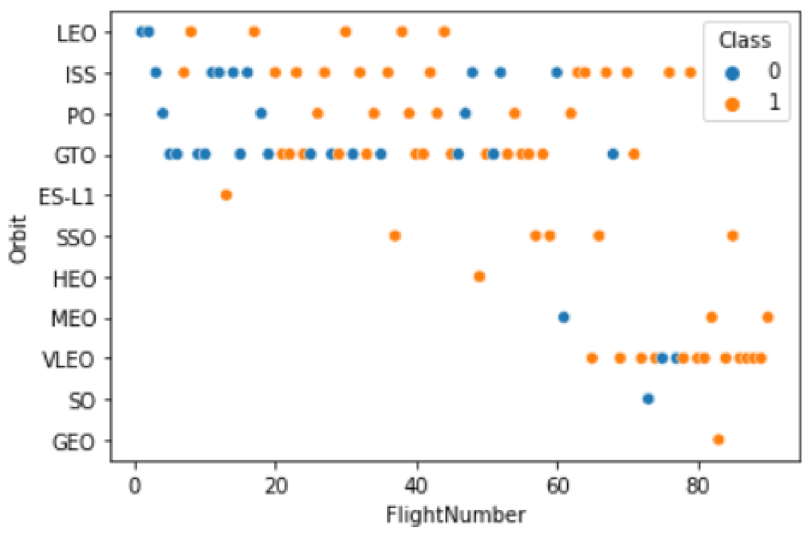
- Certain types of Orbit (e.g. GTO, ISS, LEO, MEO, PO, SO) have very low likelihood of success.

- ES-L1,GEO,SSO and HEO are the most successful – all returning a 100% success.



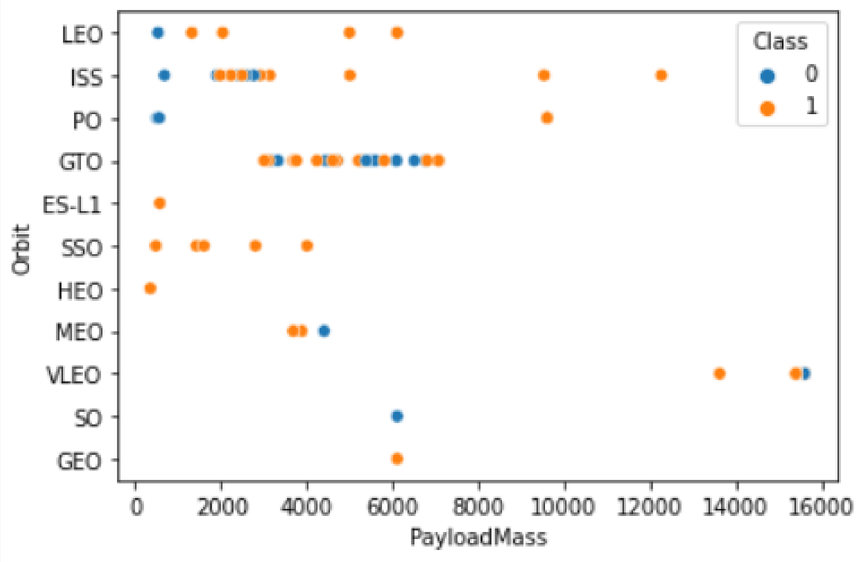
Flight Number vs. Orbit Type

- Space X initially tried out LEO, ISS, PO and GTO. These are less successful.
- As they turn to SSO and VLEO later in the development, these turn out to be more successful in landing the first stage rocket.



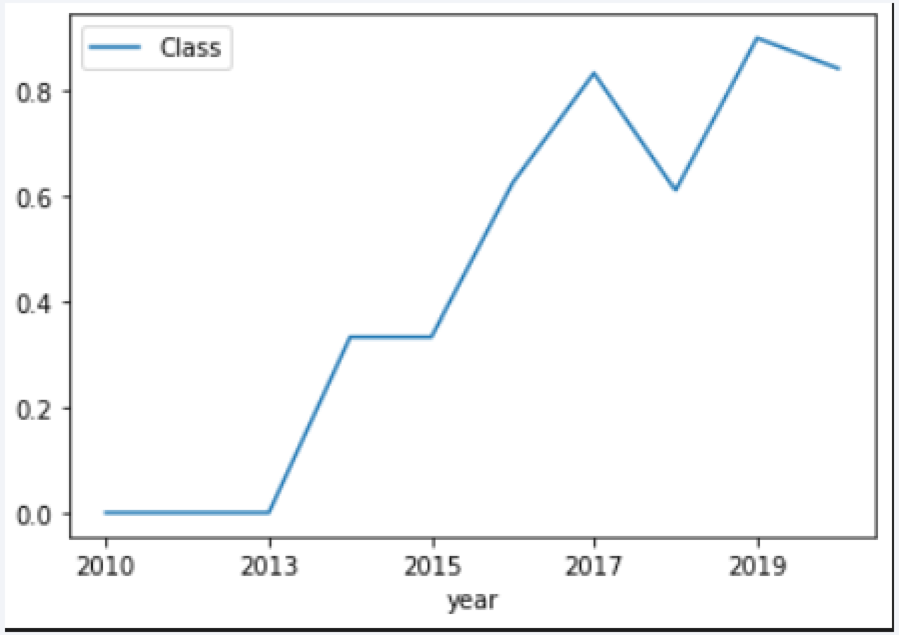
Payload vs. Orbit Type

- Many experiments were carried out for GTO in average payload weight, with mixed success
- For lighter payloads, SSO is proven to be very successful – with an almost 100% success rate.
- For heavier loads, VLEO seems to be successful as well.
- ISS seems to be able to handle a wide range of payloads, although most of the tests done were on lighter payloads (2000-4000 kg).

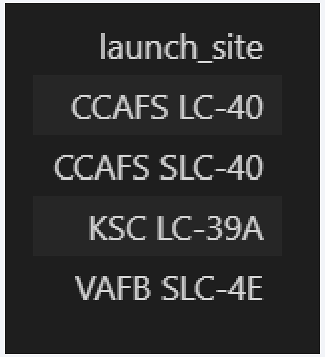


Launch Success Yearly Trend

- Success rate has drastically improved over time, by 2019 landing success peaked around 80%.



All Launch Site Names



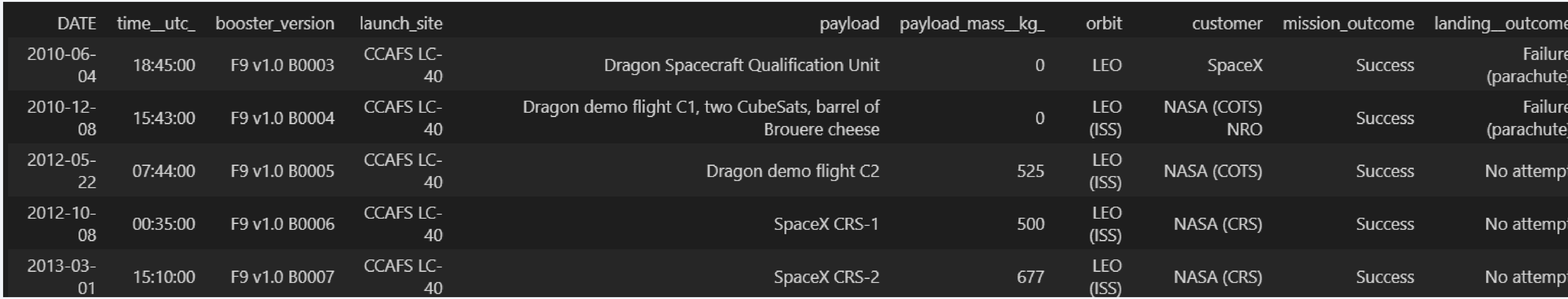
launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Generated with the following query:

```
SELECT DISTINCT(LAUNCH_SITE) FROM SPACEX
```

Which selects the distinct launch sites from the dataset.

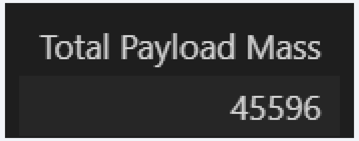
Launch Site Names Begin with 'CCA'



DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_outcome	landing_outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute
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 (parachute
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attemp
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attemp
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attemp

- Recovered with the following query:
- `SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5`
- The query selects all columns when the launch site starts with CCA, and we limit the output to the first 5 rows.

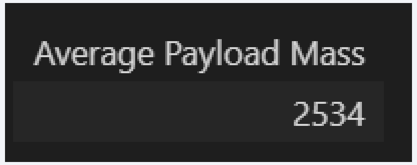
Total Payload Mass



Total Payload Mass
45596

- Recovered from the following query:
- SELECT SUM(PAYLOAD_MASS__KG_) as "Total Payload Mass" FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)’
- The query retrieves the payload column when the customer is NASA, then it performs a summation along this column and returns the sum.

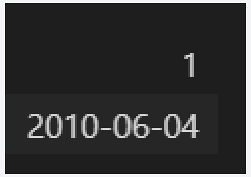
Average Payload Mass by F9 v1.1



Average Payload Mass
2534

- Recovered by the following query:
- SELECT AVG(PAYLOAD_MASS__KG_) as "Average Payload Mass" FROM SPACEX WHERE booster_version LIKE 'F9 v1.1%’
- Which takes the payload column when the booster name begins with ‘F9 v1.1’, then it takes the average along this column and returns the average value.

First Successful Ground Landing Date

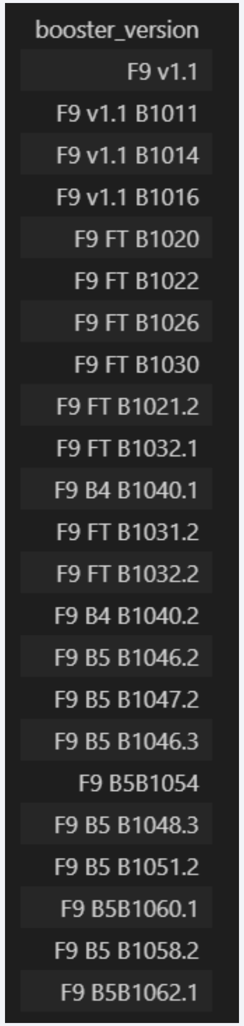


```
         1
2010-06-04
```

- Recovered by the following query:
- `SELECT MIN(DATE) FROM SPACEX WHERE MISSION_OUTCOME = 'Success’`
- It selects a subset when the mission was successful, and returns the minimum of the Date (i.e. the earliest date) in that subset.

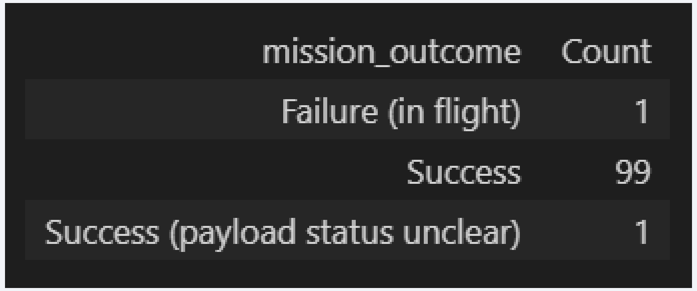
Successful Drone Ship Landing with Payload between 4000 and 6000

- Recovered by the following query:
- `SELECT booster_version FROM SPACEX WHERE mission_outcome = 'Success' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000`
- It returns the booster column from a subset of the data where the mission was successful and the payload is between 4000-6000 kg.



booster_version
F9 v1.1
F9 v1.1 B1011
F9 v1.1 B1014
F9 v1.1 B1016
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1030
F9 FT B1021.2
F9 FT B1032.1
F9 B4 B1040.1
F9 FT B1031.2
F9 FT B1032.2
F9 B4 B1040.2
F9 B5 B1046.2
F9 B5 B1047.2
F9 B5 B1046.3
F9 B5B1054
F9 B5 B1048.3
F9 B5 B1051.2
F9 B5B1060.1
F9 B5 B1058.2
F9 B5B1062.1

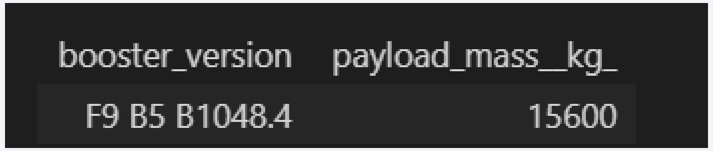
Total Number of Successful and Failure Mission Outcomes



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

- Recovered by the following query:
- `SELECT mission_outcome, COUNT(booster_version) as "Count" FROM SPACEX GROUP BY mission_outcome`
- It selects the mission outcome and groups the data by the outcome, then it counts the number of observations (here I used booster version) and returns the count by mission outcome.

Boosters Carried Maximum Payload



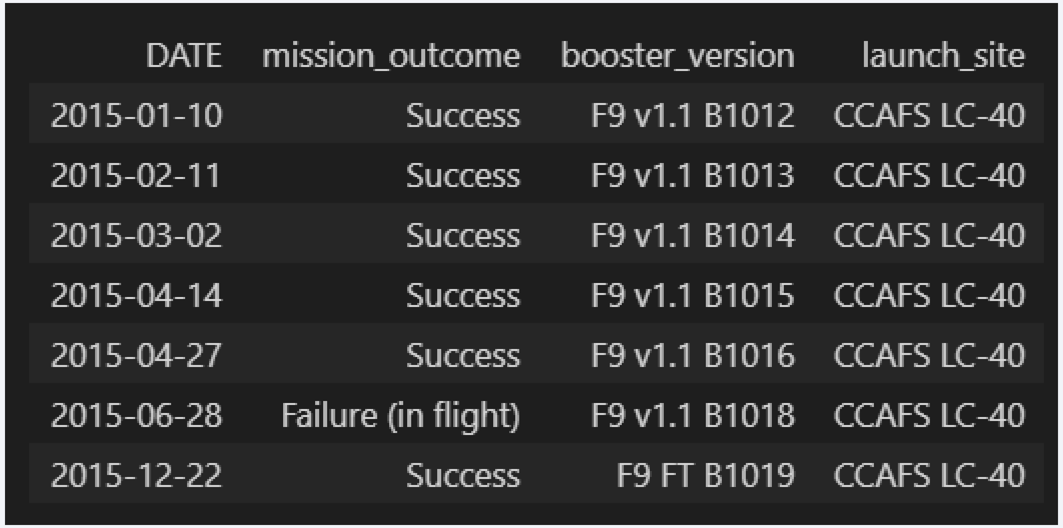
booster_version	payload_mass__kg_
F9 B5 B1048.4	15600

- Recovered by the following query:

- SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEX ORDER BY PAYLOAD_MASS__KG_ DESC LIMIT 1

- It sorts the dataset by payload mass in the descending order (so the row with the heaviest payload is on top), then it returns first the booster version and the payload – i.e. the record with the max payload.

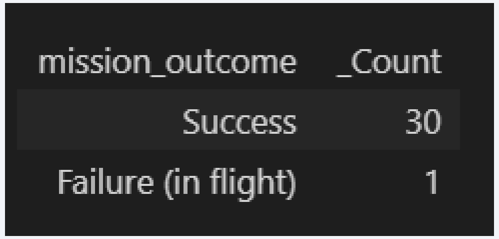
2015 Launch Records



DATE	mission_outcome	booster_version	launch_site
2015-01-10	Success	F9 v1.1 B1012	CCAFS LC-40
2015-02-11	Success	F9 v1.1 B1013	CCAFS LC-40
2015-03-02	Success	F9 v1.1 B1014	CCAFS LC-40
2015-04-14	Success	F9 v1.1 B1015	CCAFS LC-40
2015-04-27	Success	F9 v1.1 B1016	CCAFS LC-40
2015-06-28	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
2015-12-22	Success	F9 FT B1019	CCAFS LC-40

- Recovered by the following query:
- SELECT DATE, MISSION_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE DATE LIKE '2015%’
- It selects the date, mission outcome, booster version and launch site when the date begins with the year 2015.

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



mission_outcome	_Count
Success	30
Failure (in flight)	1

- Recovered by the following query:

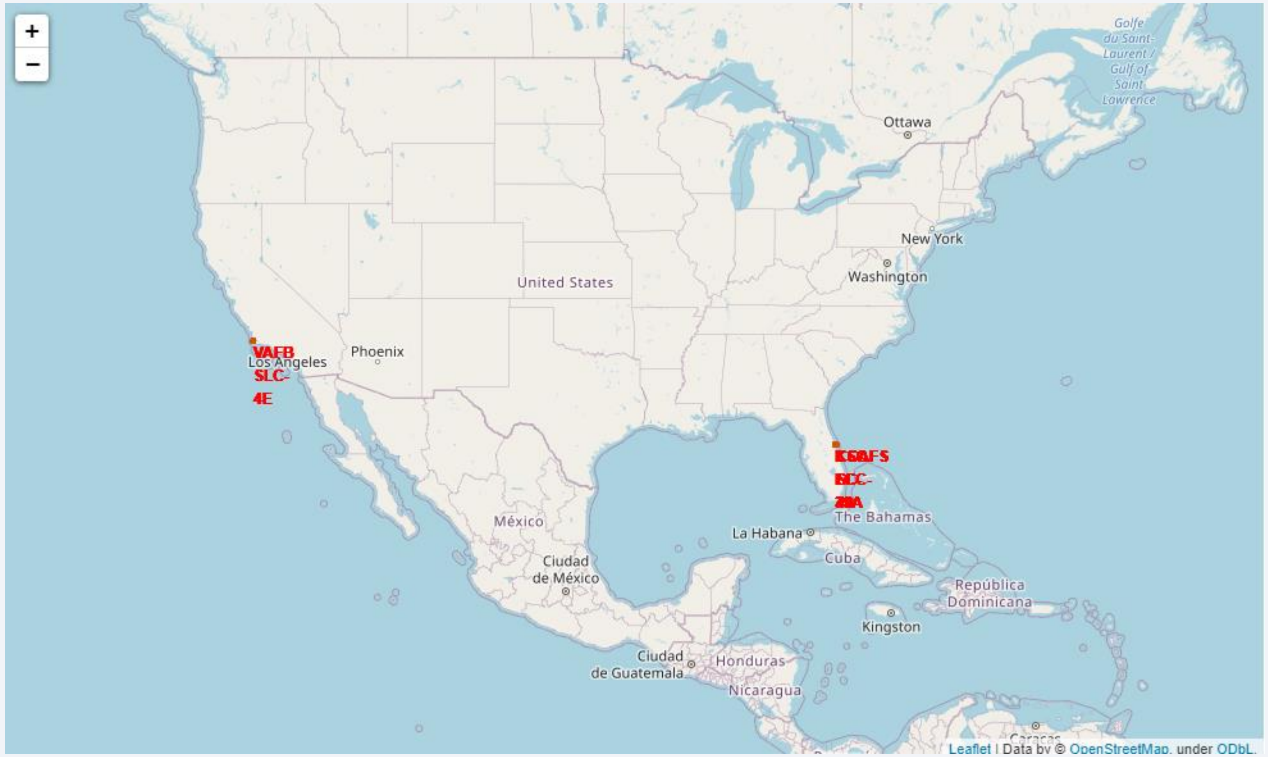
-
  ```
  SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) AS "_Count" FROM SPACEX WHERE DATE
  BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY MISSION_OUTCOME ORDER BY "_Count"
  DESC
  ```

- It first subsets the data such that the date is between 2010/6/4 and 2017/3/20, then it groups the rows by mission outcome. It then counts the observations by their mission outcome, and order the output by the count, so we end up with the most likely outcome on top (i.e. success)

Section 3

Launch Sites Proximities Analysis

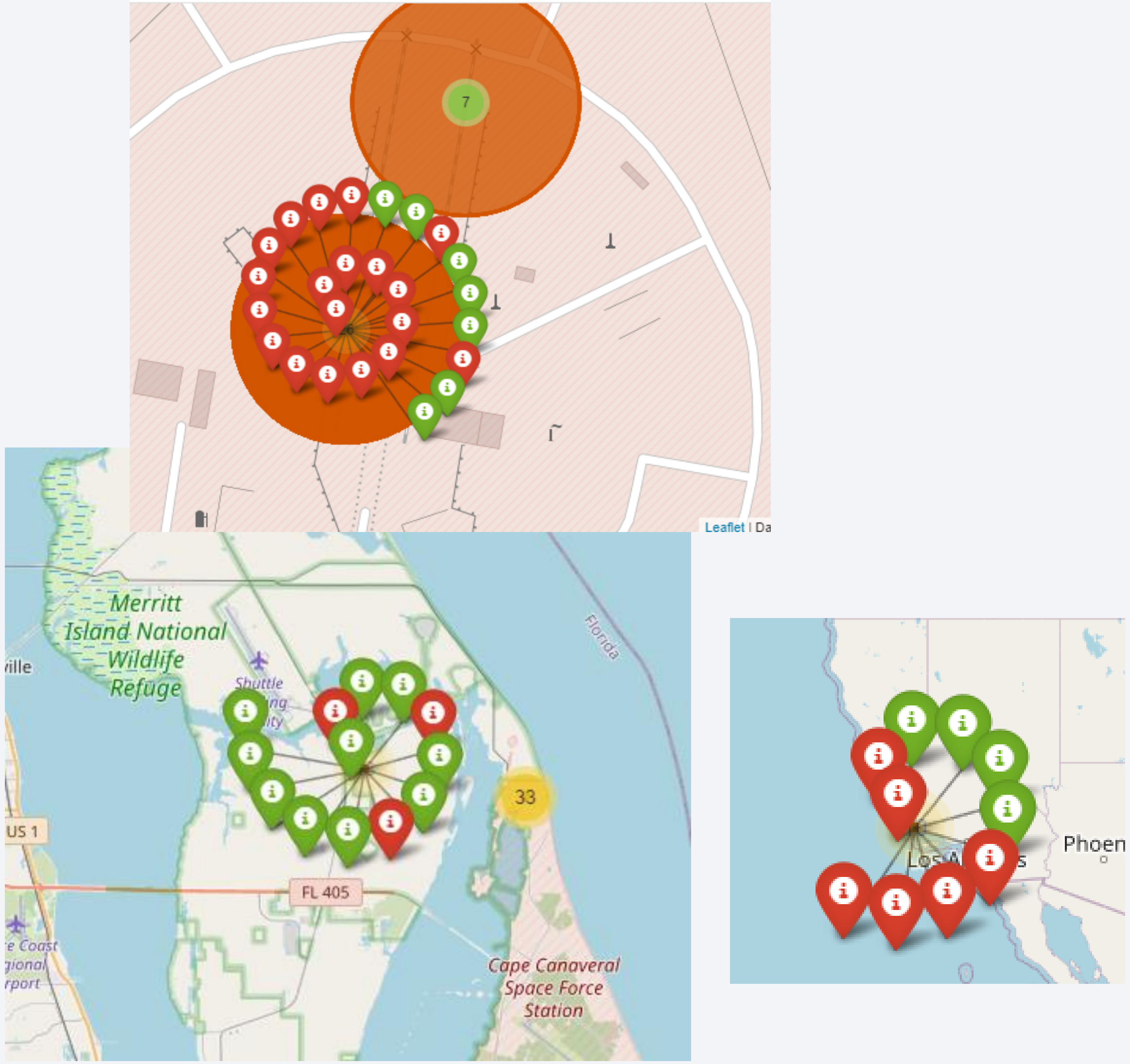
Location of launch sites



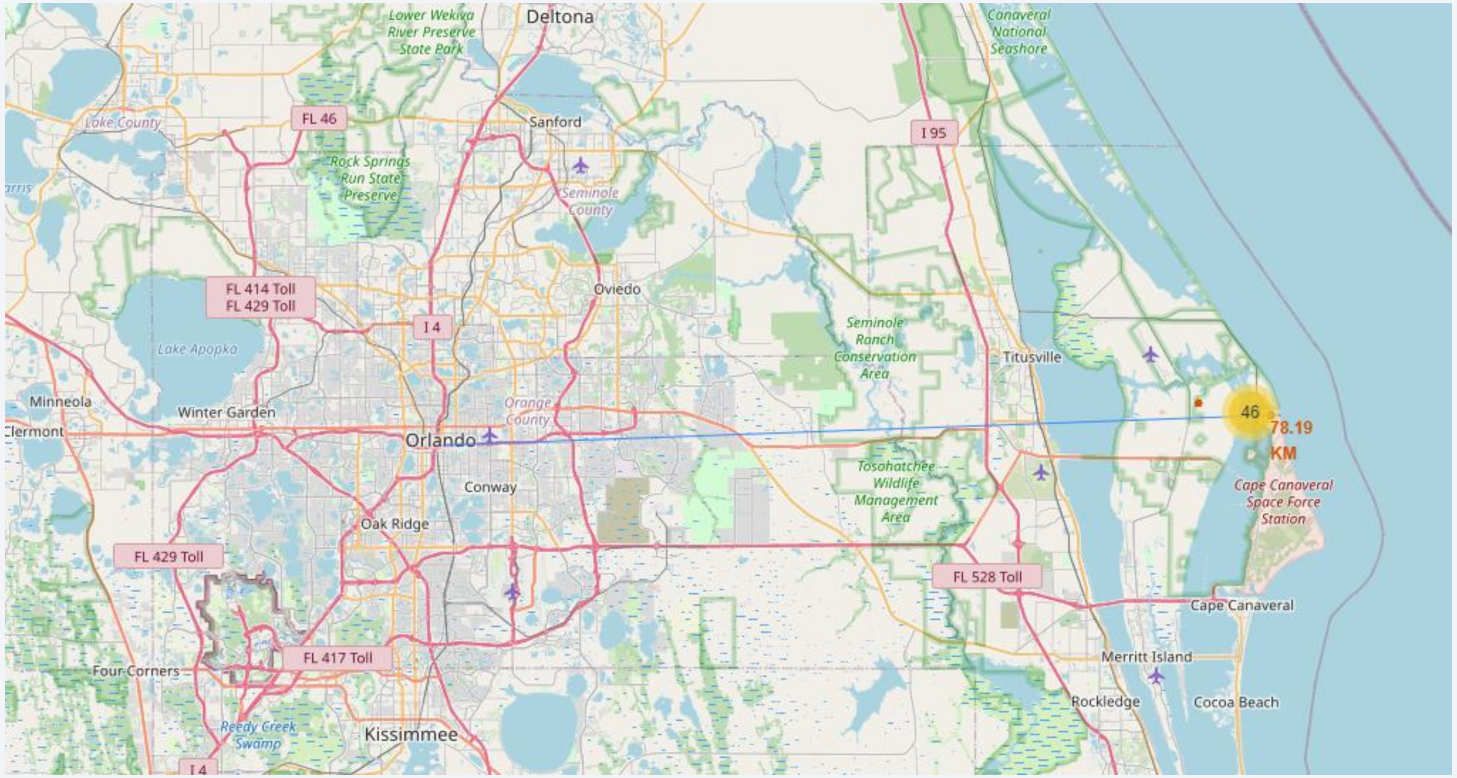
SpaceX has launched the Falcon 9 tests from both Southern California and Western Florida. The majority of the tests were carried out in Florida.

Where do the successful landings come from?

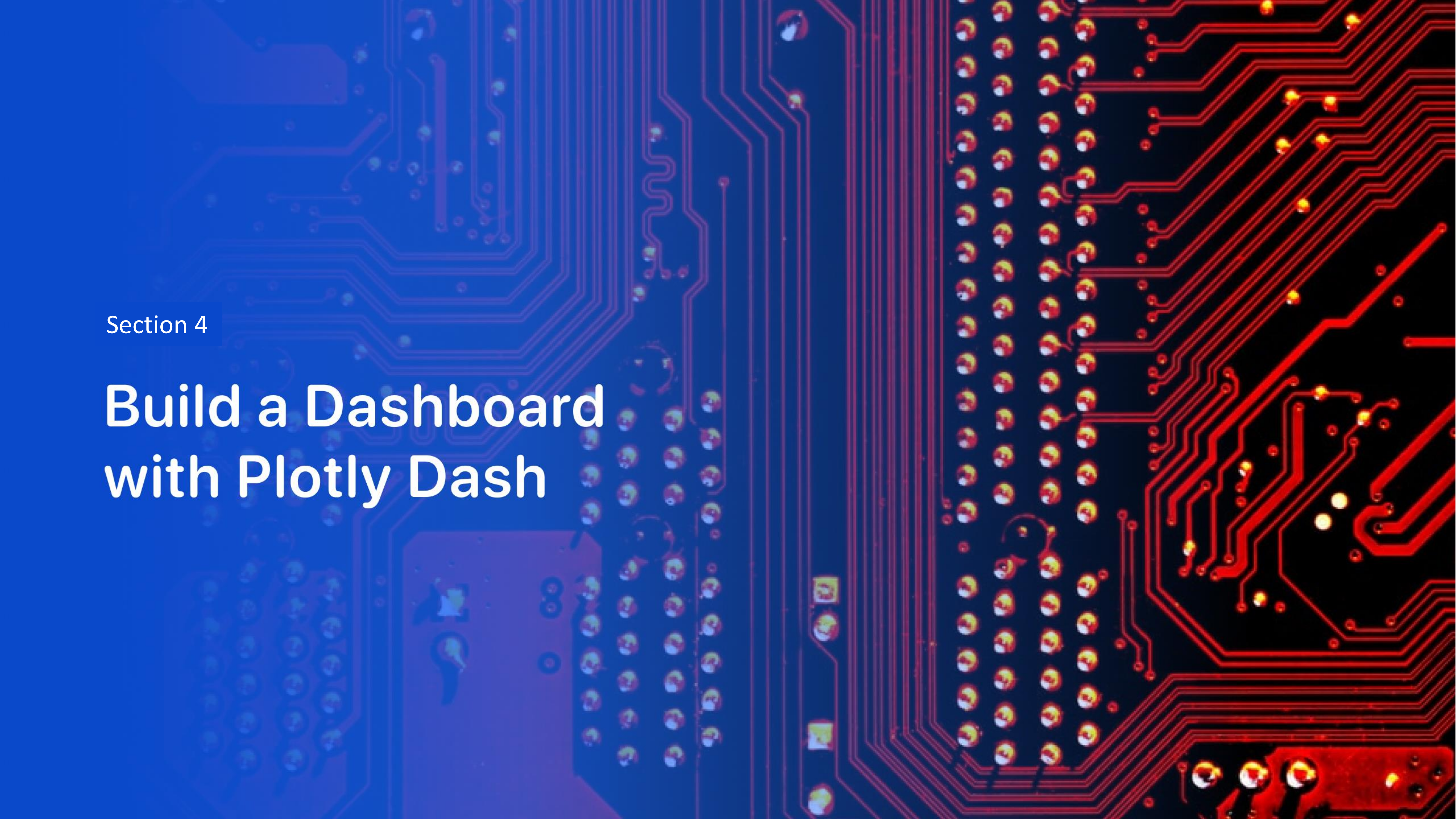
- A majority of the failed missions were launched in the site that is closest to the shoreline.
- The in-land site is more successful in landing the first-stage rocket.
- The Southern California site has a rather mixed result in mission outcomes.



Proximity of the Florida Launch sites to the nearest major city



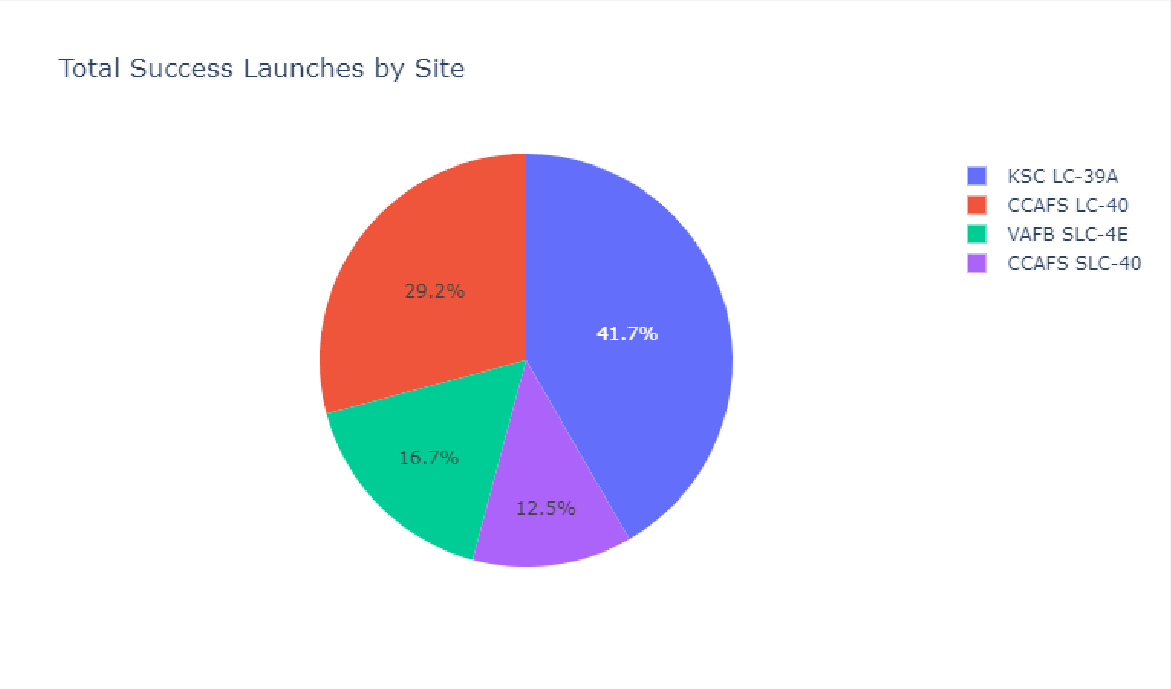
- Although the Florida launch sites are on the shoreline, it is still somewhat close to Orlando, the closest major city. It is only 78 km away.
- In case of a major accident, debris might fall into the city limits of Orlando. Therefore being able to maximize the likelihood of landing success is critical!



Section 4

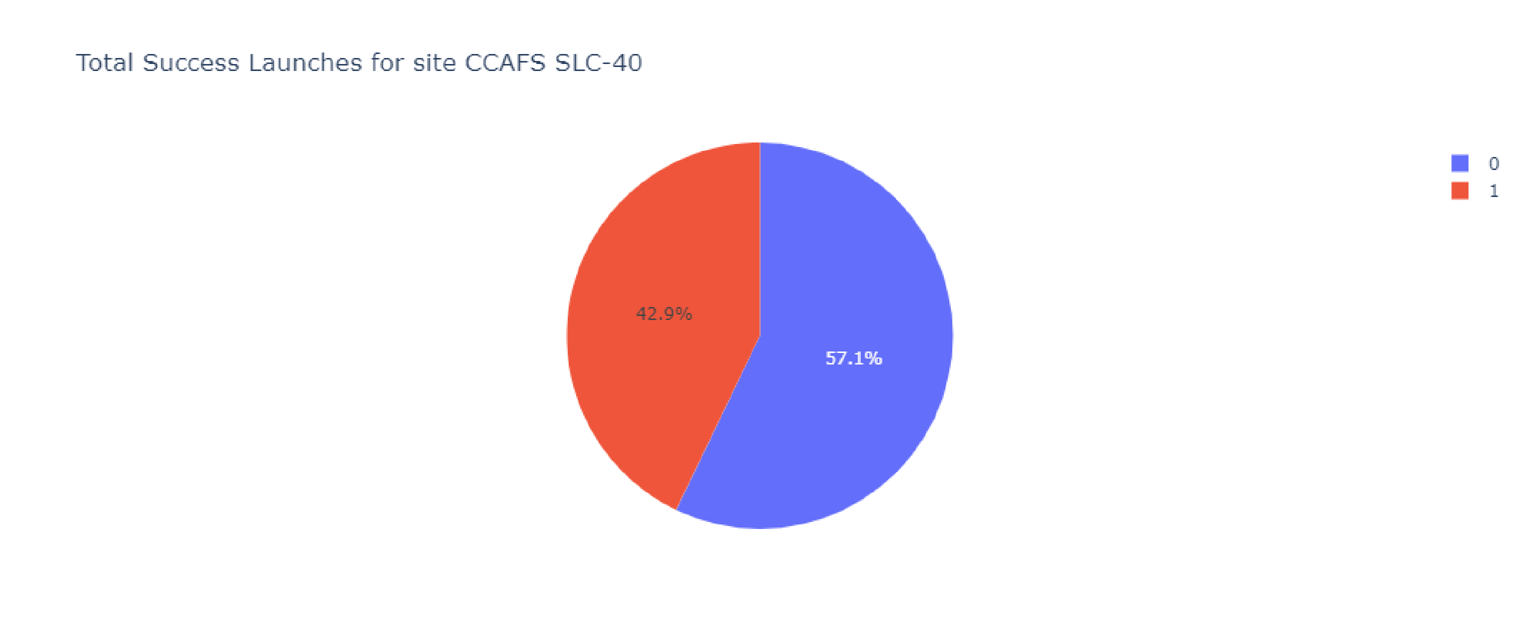
Build a Dashboard with Plotly Dash

Most successful launch site



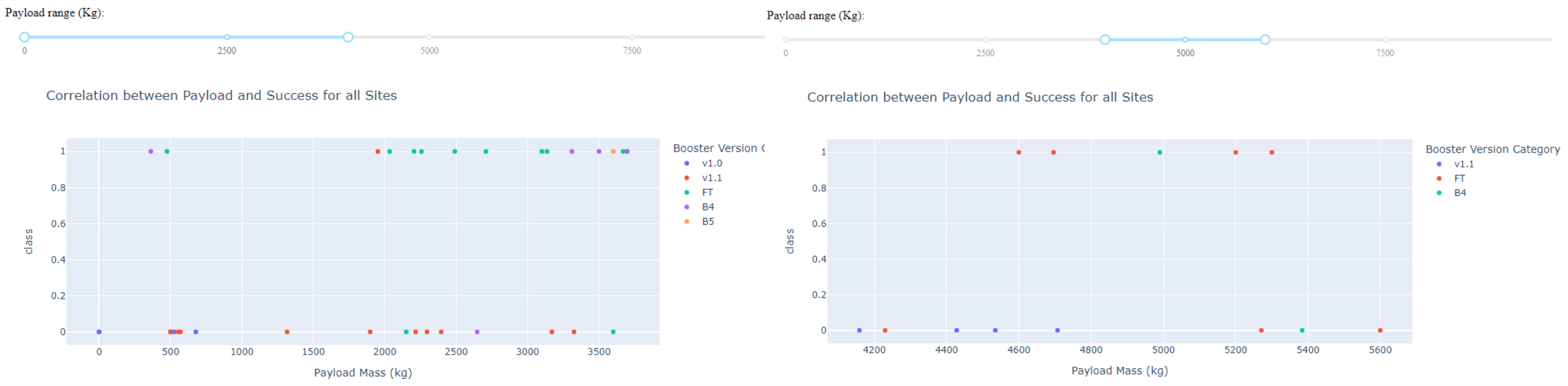
- The most successful launch site is KSC LC-39A.

The most successful launch site by success-to-fail ratio



- Some launch sites launched more tests than others, so a success-to-failure ratio is fairer.
- CCAFS SLC-40 is the most successful using this metric, with a 42.9% success rate.

Lighter payloads are more likely to succeed

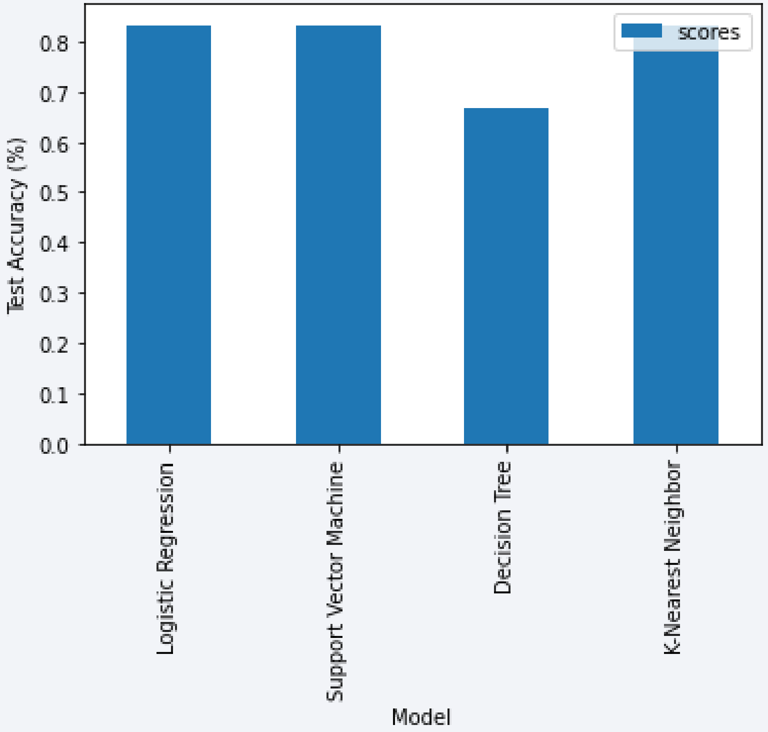


- Overall, flights with lighter payloads are more likely to succeed.
- As the payload get to an average weight between 4000-6000 kg, the success rate drops drastically.

Section 5

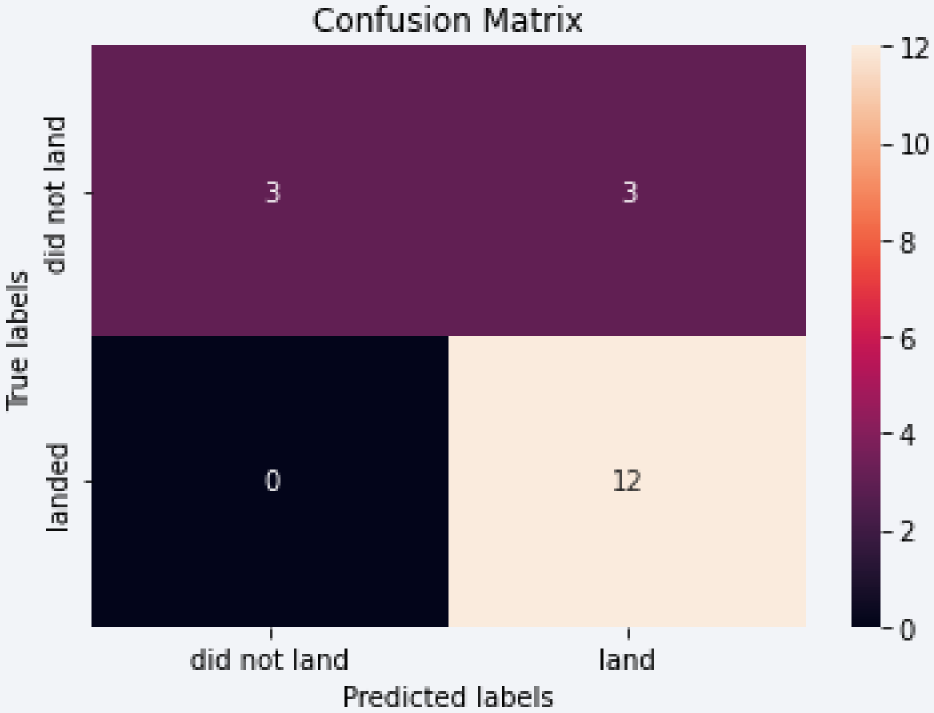
Predictive Analysis (Classification)

Classification Accuracy



- The Logistic Regression, Support Vector Machine and K-Nearest Neighbor models performed better than the Decision Tree. They all have an accuracy score of around 83.33%.

Confusion Matrix



- This confusion matrix is taken from the logistic regression model.
- It shows that the model was able to predict successful outcomes to be successful, and vice versa.
- There's a slight problem with false positives – the model predicted 3 cases to be successful whereas they actually failed.

Conclusions

- Falcon-9 has come a long way since 2010, the latest tests are more likely to be successful, yielding a 80% success rate.
  - As SpaceX continues to refine the reliability of the first-stage rocket, Falcon-9 has a huge potential in bringing down the cost of space travel and space exploration.
- Payload mass remains the main factor in determining mission success. A lighter payload is always more likely to succeed.
  - When deploying Falcon-9, there should be a payload mass restriction such that the first-stage rocket is less likely to fail -> customer less likely have to pay to acquire a new one
- Certain types of Orbit are more likely to succeed.
  - For lighter payloads, SSO is proven to be very successful – with an almost 100% success rate.
  - For heavier loads, VLEO seems to be successful as well.
  - ISS seems to be able to handle a wide range of payloads, although most of the tests done were on lighter payloads (2000-4000 kg).

Appendix

- Code to visualize model performance in a bar chart is available here:

```
names = ["Logistic Regression", "Support Vector Machine", "Decision Tree", "K-Nearest Neighbor"]
scores = [logreg_cv.score(X_test, Y_test),
            svm_cv.score(X_test, Y_test),
            tree_cv.score(X_test, Y_test),
            knn_cv.score(X_test, Y_test)]

plotdf = pd.DataFrame({'model':names, 'scores':scores})
plotdf.plot(kind="bar", x="model", y="scores")
plt.xlabel("Model")
plt.ylabel("Test Accuracy (%)")
```

Thank you!