Winning Space Race with Data Science

Errikos Kiladis
28-11-2022

Outline

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

Executive Summary

- Summary of methodologies
  Data collection, data wrangling, exploratory data analysis (EDA), data visualization, model development, model evaluation, and reporting results to stakeholders.

- Summary of all results
  1. Defined and formulated a real-world business problem to find out interesting insights.
  2. Extract meaningful patterns by using data visualization.
  3. Development of sophisticated tools such as Interactive Visual Analytics and Dashboards.
  4. Optimization of the predictive model regarding test data set.
  5. Delivery and presentation of the data-driven insights to determine if the first stage of Falcon 9 will land successfully.

Introduction

- Project background and context
  By assuming the role of a Data Scientist working for a startup intending to compete with SpaceX, we are asked to predict if the Falcon 9 first stage will land successfully. SpaceX 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 SpaceX can reuse the first stage.
- Problems you want to find answers
  We want to determine if the first stage will land in order to determine the cost of a launch. This information can be used by our startup company who wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API
  - Web Scraping from Wikipedia
- Perform data wrangling
  - Perform exploratory Data Analysis and determine Training Labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build, tune, and evaluate classification models

Data Collection

- Data collection was implemented using two different sources, namely, SpaceX API and Wikipedia.

Data Sources

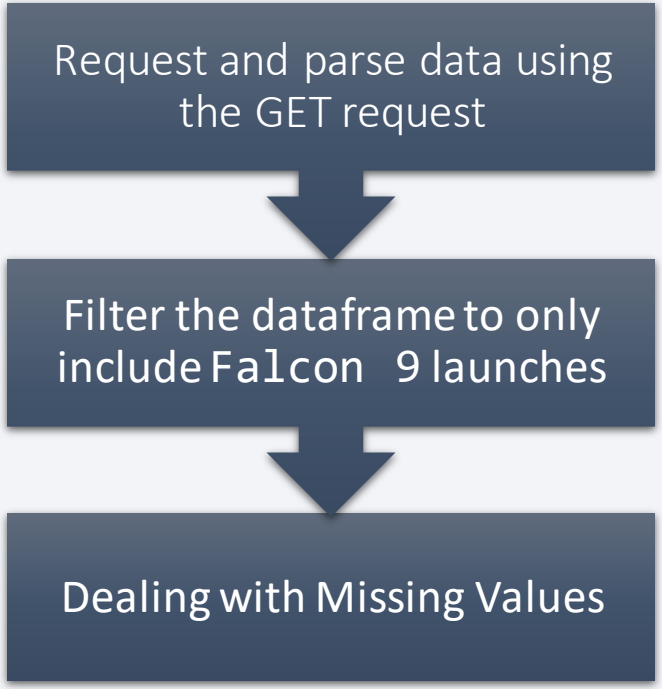
- SpaceX API: https://api.spacexdata.com/v4/launches/past
- Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

Objectives:

- Request to the SpaceX API
- Clean the requested data

Click the link bellow to open the notebook:
Data Collection API

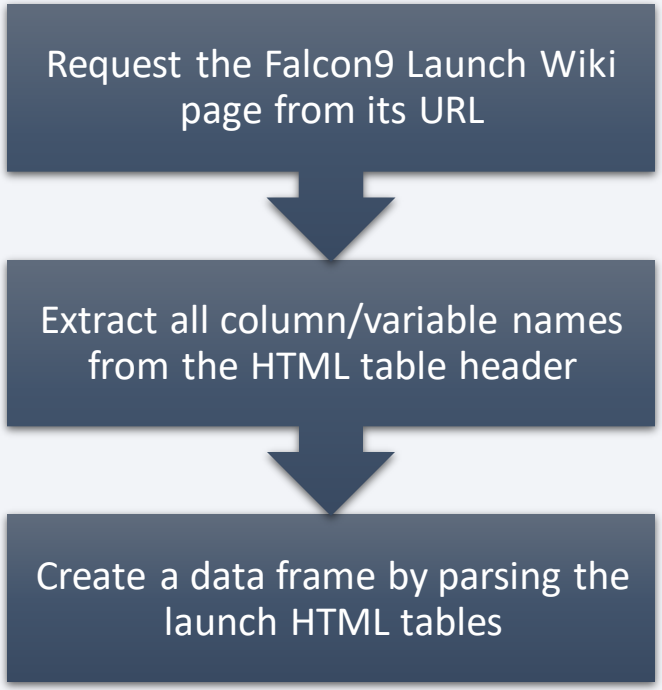


Data Collection - Scraping

Objectives:

- Web scrap Falcon 9 launch records with BeautifulSoup
- Extract a Falcon 9 launch records HTML table from Wikipedia
- Parse the table and convert it into a Pandas data frame

Click the link bellow to open the notebook:
Data Collection Web Scraping



Data Wrangling

- Load SpaceX dataset
- Identify and calculate the percentage of the missing values in each attribute
- Identify which columns are numerical and categorical

Click the link bellow to open the notebook:
Data Wrangling

EDA with Data Visualization

- Using basic visualization tools such as ***bar charts***, ***scatter point charts***, and ***line charts***, helped as to evaluate the significance of a variety of features and their relationship between them and most importantly their relationship with the outcome of the landing.

Click the link bellow to open the notebook:
EDA using Pandas & Matplotlib

EDA with SQL

SQL queries:

- The names of the unique launch sites in the space mission
- 5 records where launch sites begin with the string 'CCA'
- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1
- The date when the first successful landing outcome in ground pad was achieved
- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- The total number of successful and failure mission outcomes
- The names of the booster_versions which have carried the maximum payload mass. Use a subquery
- 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 the date 2010-06-04 and 2017-03-20, in descending order

Click the link bellow to open the notebook:

EDA using SQL

Build an Interactive Map with Folium

In order to provide powerful visualization features, we created and added to the folium maps the following objects:

- Map: to initialize a world map representation
- Circle: to add a highlighted circle area with a text label on a specific coordinate
- Marker: to easily identify which launch sites have relatively high success rates

Click the link bellow to open the notebook:
Interactive Visual Analytics with Folium

Build a Dashboard with Plotly Dash

- The dashboard application that was developed, contains a pie chart and a scatter point chart which are manipulated by dropdown and range slider functionalities.
- We added the pie chart and the scatter point chart because we want to find if variable payload is correlated to mission outcome. From a dashboard point of view, we want to be able to easily select different payload range and see if we can identify some visual patterns.

Click the link bellow to open the notebook:
SpaceX Dashboard Application

Predictive Analysis (Classification)

- Performed EDA and determined Training Labels
- Find best Hyperparameters for Logistic Regression, SVM, Classification Trees, and KNN models using GridSearchCV
- Evaluate the optimum models of each method by checking the test accuracy and monitoring the confusion matrices
- Choose the best among the optimum models considering the test accuracy criterion as critical

Click the link bellow to open the notebook:
Machine Learning Prediction

Results

Exploratory data analysis results

- Different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000).
- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- The success rate since 2013 kept increasing till 2020.

Results

- Interactive analytics
- Launch sites are near by the sea and have facilitated access to critical infrastructures such as airports, train stations, ports, highways, and big cities.
- Obviously KSC LC-39A site has the highest success rate while CCAFS LC-40 is the most used launch site.

Results

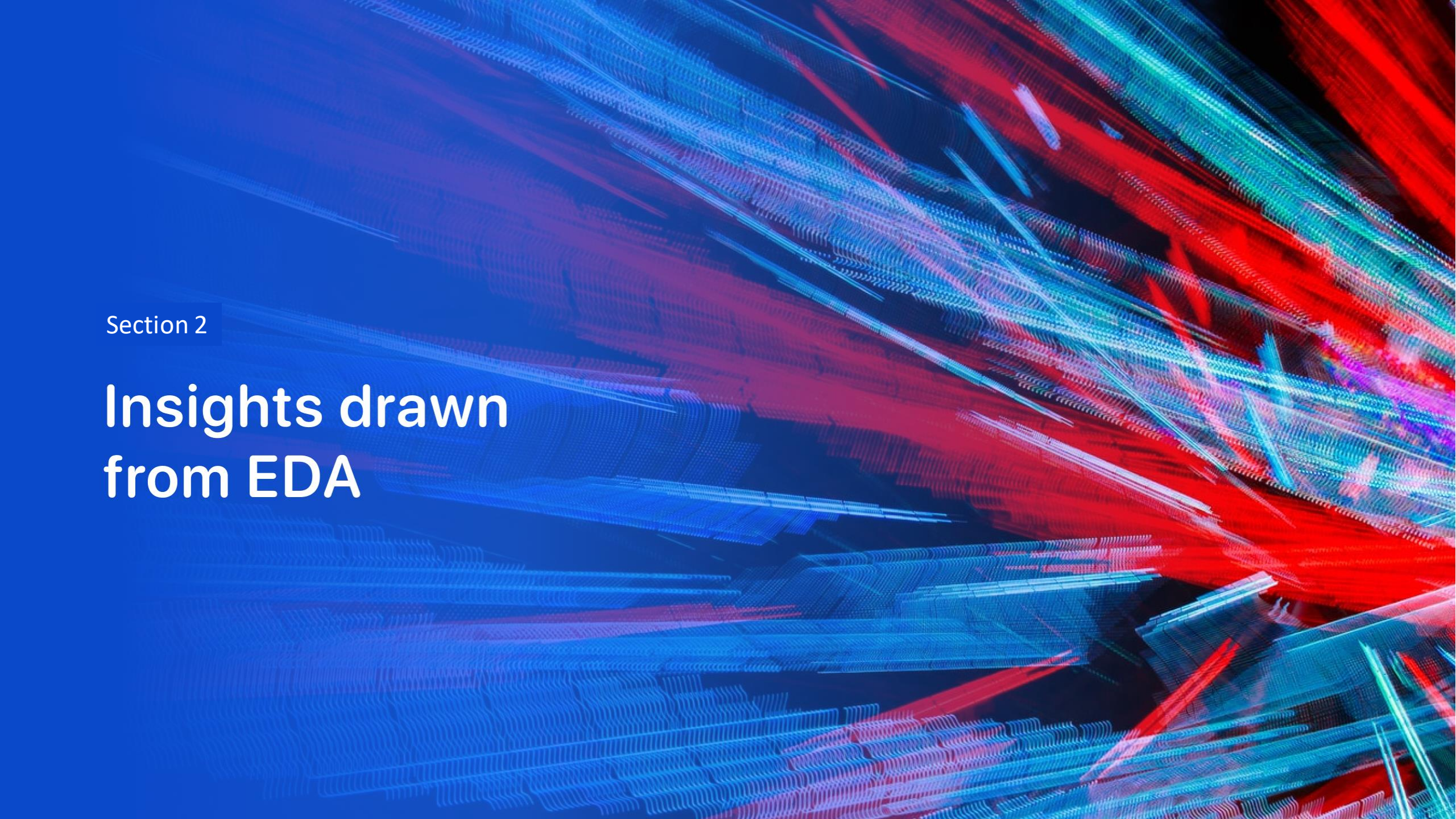
- Predictive analysis results

As we can mention from the table below, the model with the best performance in terms of test accuracy was "Decision Tree Classifier" which scored 0.9444 on the test data set and 0.875 on the training set.

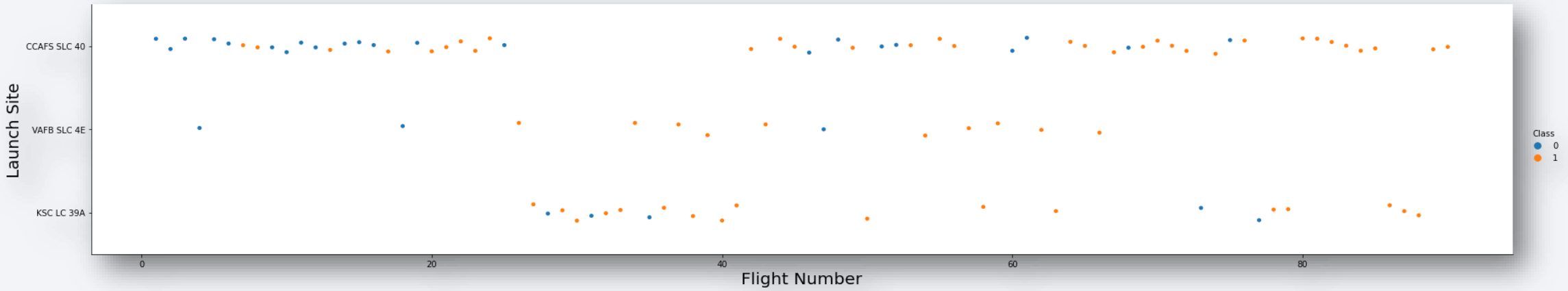
Model	Accuracy	Test Accuracy
Logistic Regression	0.8464	0.8333
Support Vector Machine	0.8482	0.8333
Decision Tree Classifier	**0.875**	**0.9444**
K Nearest Neighbors	0.8482	0.8333

Section 2

Insights drawn from EDA



Flight Number vs. Launch Site



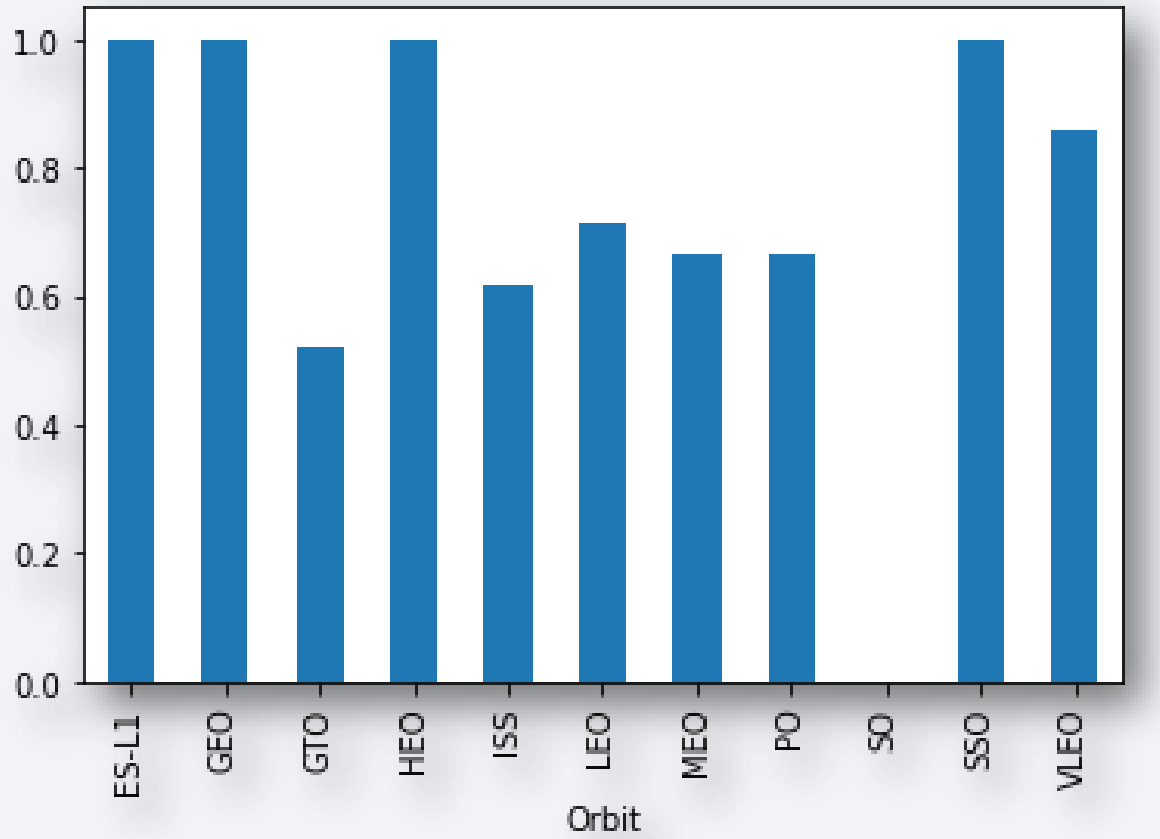
- By observing carefully the scatter point chart above, we notice that the vast majority of the first 20 flights did not land successfully. Also, it is clear that most of the latest flights landed successfully which is very reasonable, considering the experience that is gained through the iteration of the process.

Payload vs. Launch Site

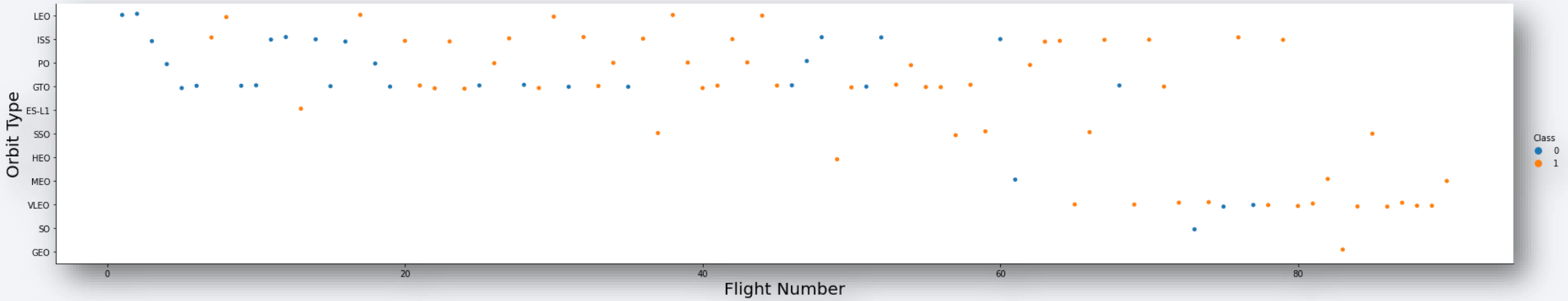
- A very exciting finding that we extract from the Payload vs. Launch Site chart, is that there is only one failure among the last trials which were conducted with a payload bigger than 12000 kg. In other words, higher success rates comes with heavier payload. Also, for the VAFB-SLC launch site there are no rockets launched for heavy payload mass greater than 10000 kg.

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO orbits have success rates equal to 1. VLEO is following, while the rest of the orbits have success rates between 0.5 and 0.7, except SO orbit which has success rate equal to 0.

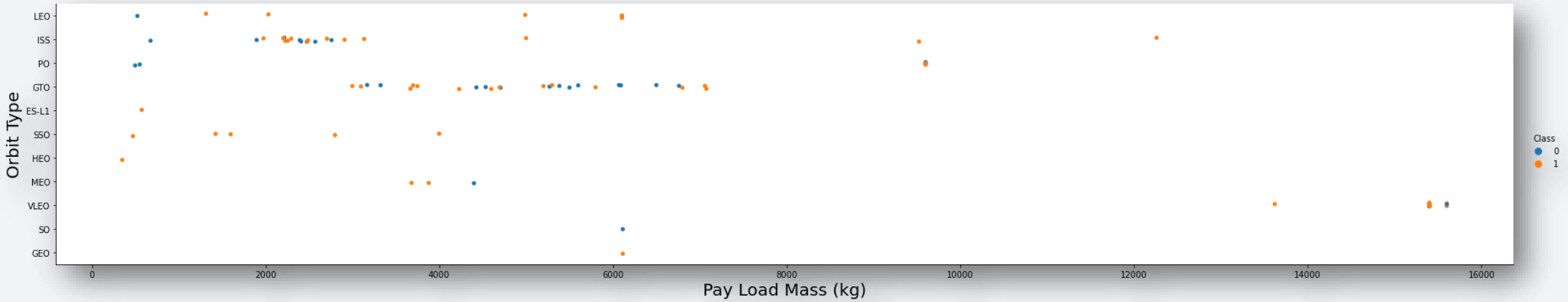


Flight Number vs. Orbit Type



- In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

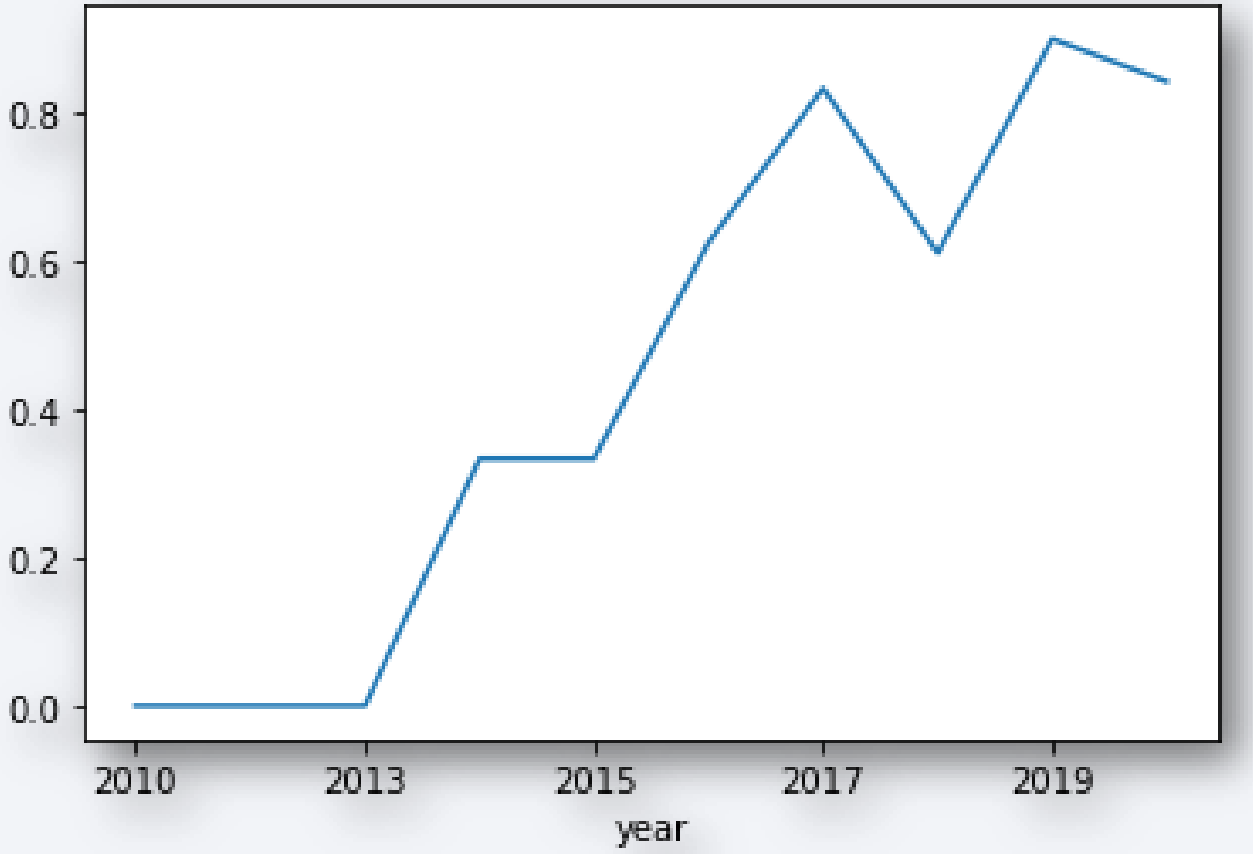
Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for PO, LEO and ISS. However, for GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are kind of mixed.

Launch Success Yearly Trend

- It is quite clear that the success rate since 2013 kept increasing till 2020. In 2018 we observe a significant decrease. In order to find out the reasons that caused this kind of decreasing, we should further investigate our data, especially for the years 2017 to 2020.



All Launch Site Names

- By using the **DINSTINCT** statement we detect the unique launch sites from the SpaceX data set.

Display the names of the unique launch sites in the space mission

```
%%sql

SELECT DISTINCT LAUNCH_SITE FROM SPACEX;
```

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

Launch Site Names Begin with 'CCA'

- By using the **LIKE** statement (CCA%) we grab only the rows where the launch site begins with the letters CCA. Also, we use **LIMIT 5** to present only the first 5 rows of the query results.

```
%%sql

SELECT * FROM SPACEX
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

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 attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The query below calculates the **SUM** of the payload mass in kg for all the missions of a specific customer, namely, NASA (CRS).

```
%%sql

SELECT SUM(PAYLOAD_MASS__KG_) AS total_payload_mass_carried_by_boosters_launched_by_NASA_in_kg FROM SPACEX
WHERE CUSTOMER = 'NASA (CRS)';
```

total_payload_mass_carried_by_boosters_launched_by_nasa_in_kg
45596

Average Payload Mass by F9 v1.1

- In order to calculate the average payload mass by F9 v1.1 we use the **AVG** function and the LIKE statement to include only the F9 V1.1 payloads in our calculation.

```
%%sql

SELECT AVG(PAYLOAD_MASS__KG_) AS average_payload_mass_carried_by_booster_version_F9_v1_1_in_kg
FROM SPACEX
WHERE BOOSTER_VERSION LIKE 'F9 v1.1';
```

average_payload_mass_carried_by_booster_version_f9_v1_1_in_kg
2928

First Successful Ground Landing Date

- From the SpaceX dataset, we extract only the rows where the landing outcome is equal to 'Success (ground pad)'. Then we grab the minimum date which is actually the first successful ground landing date.

```
%%sql

SELECT MIN(DATE)
AS DATE_WHEN_THE_FIRST_SUCCESSFUL_LANDING_OUTCOME_IN_GROUND_PAD_WAS_ACHIEVED
FROM SPACEX
WHERE LANDING__OUTCOME LIKE 'Success (ground pad)';
```

date_when_the_first_successful_landing_outcome_in_ground_pad_was_achieved
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The condition which was used to grab the rows has three different parts which have to be true.

1. LANDING_OUTCOME = 'Success (drone ship)'
2. Payload_mass_kg_<6000
3. Payload_mass_kg_>4000

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

```
%%sql

SELECT BOOSTER_VERSION AS LIST_OF_BOOSTERS_NAMES
FROM SPACEX
WHERE LANDING__OUTCOME = 'Success (drone ship)' AND payload_mass__kg_<6000 AND payload_mass__kg_>4000;
```

Total Number of Successful and Failure Mission Outcomes

- In order to calculate the total number of successful and failure mission outcomes, we grouped by the mission outcomes and then we cummulated all the rows of each possible outcome using the COUNT statement.

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

```
%%sql

SELECT MISSION_OUTCOME, COUNT(*) AS TOTAL_NUMBER FROM SPACEX GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

Boosters Carried Maximum Payload

- Here we used a nested query to grab the booster versions with a payload mass equal to the maximum payload mass from the SpaceX dataset.

```
%%sql

SELECT BOOSTER_VERSION AS NAMES_OF_THE_BOOSTER_VERSIONS from SPACEX
WHERE payload_mass__kg_ = (SELECT MAX(payload_mass__kg_) FROM SPACEX);
```

names_of_the_booster_versions
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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 were grabbed by using two conditions which both have to be true.

1. LANDING_OUTCOME = 'Failure (drone ship)'
2. The year of the landing is 2015. [year(date)=2015]

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

```
%%sql

SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX where LANDING__OUTCOME = 'Failure (drone ship)' AND year(date)=2015;
```

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

- First we grouped by landing outcome and we ordered it by count. Then we asked for the count of all the rows with a date between 2010-06-04 and 2017-03-20, in descending order.

```
%%sql

SELECT LANDING__OUTCOME,COUNT(*) AS COUNT_ FROM SPACEX

WHERE DATE BETWEEN '2010-06-04' and '2017-03-20'

GROUP BY LANDING__OUTCOME ORDER BY COUNT_ DESC;
```

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

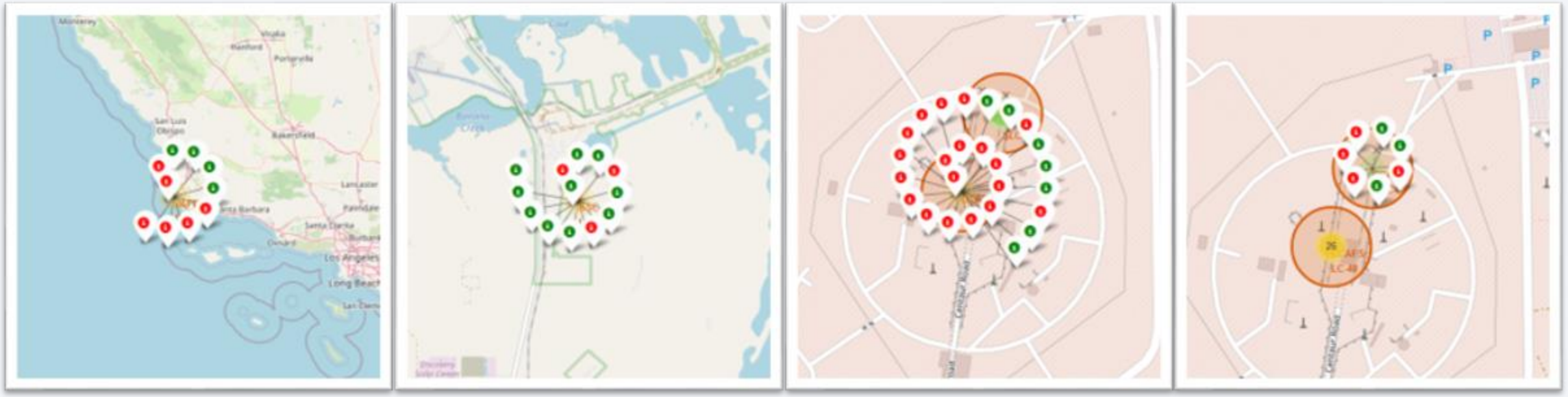

Section 3
Launch Sites
Proximities Analysis

Launch sites on a global map

- VAFB SLC-4E is the only launch site placed in the west coast.
- The rest 3 launch sites are placed in the east coast.

Success-Failed launches for each site on the map

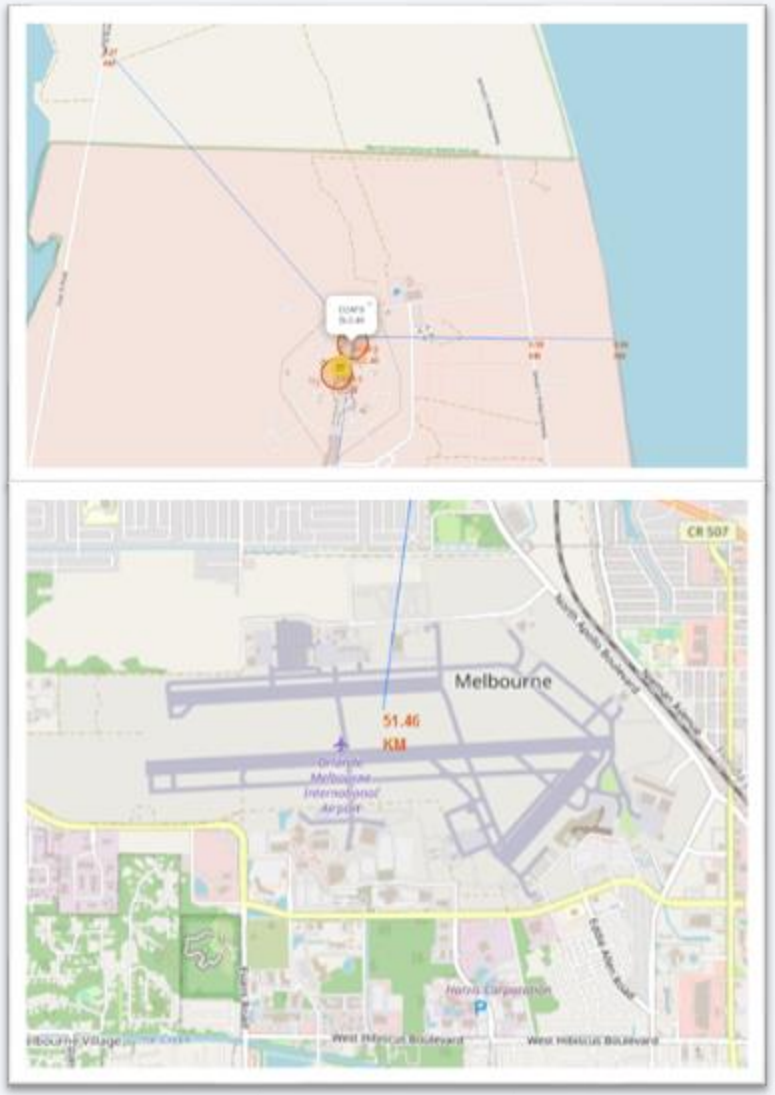
- The visualization of the mission outcome on the map, provides an extra ordinary opportunity to have a quick but also accurate understanding of the success rate for each of the launch sites.

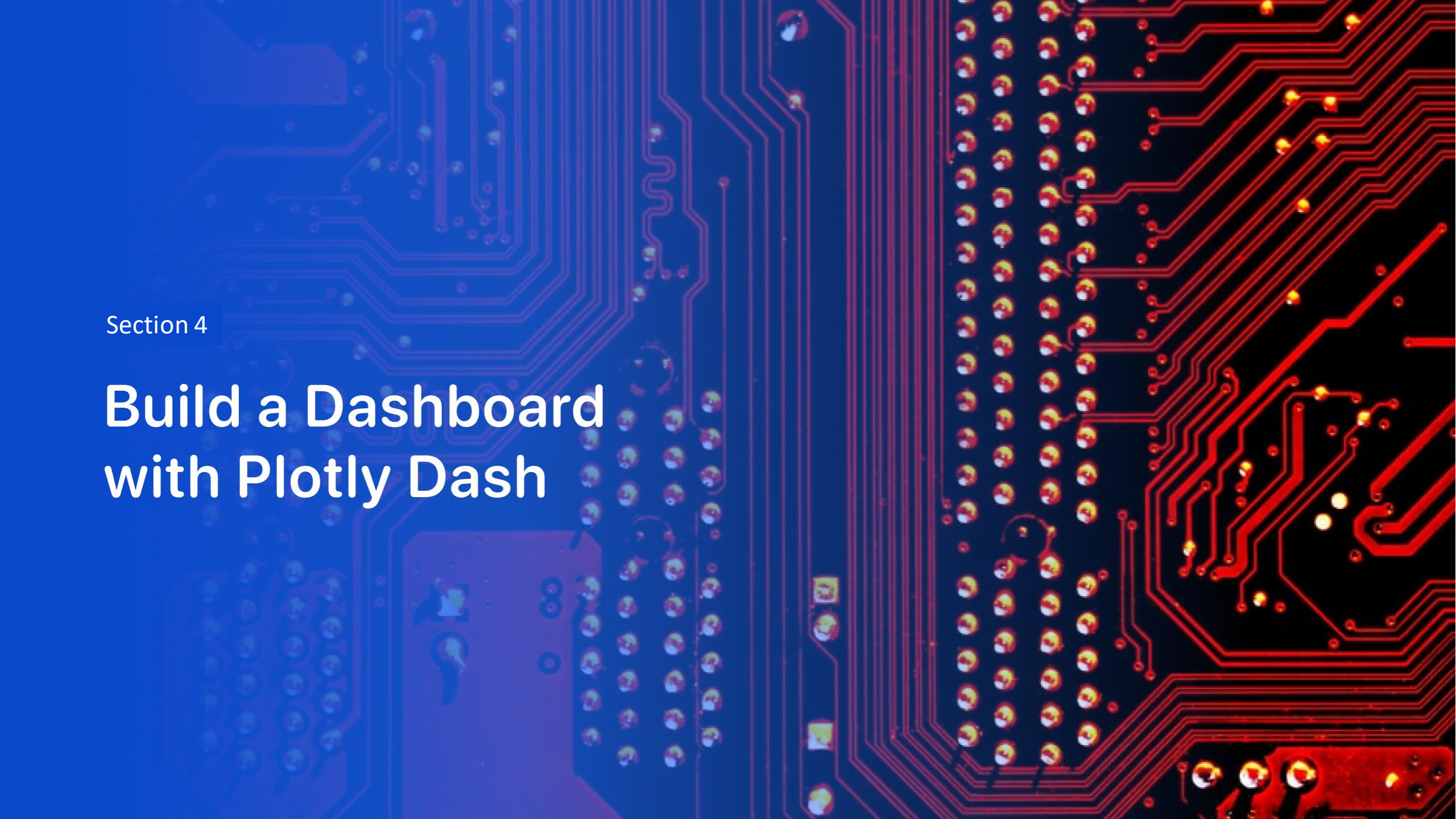


Proximity to critical infrastructures

CCAFS SLC-40 is

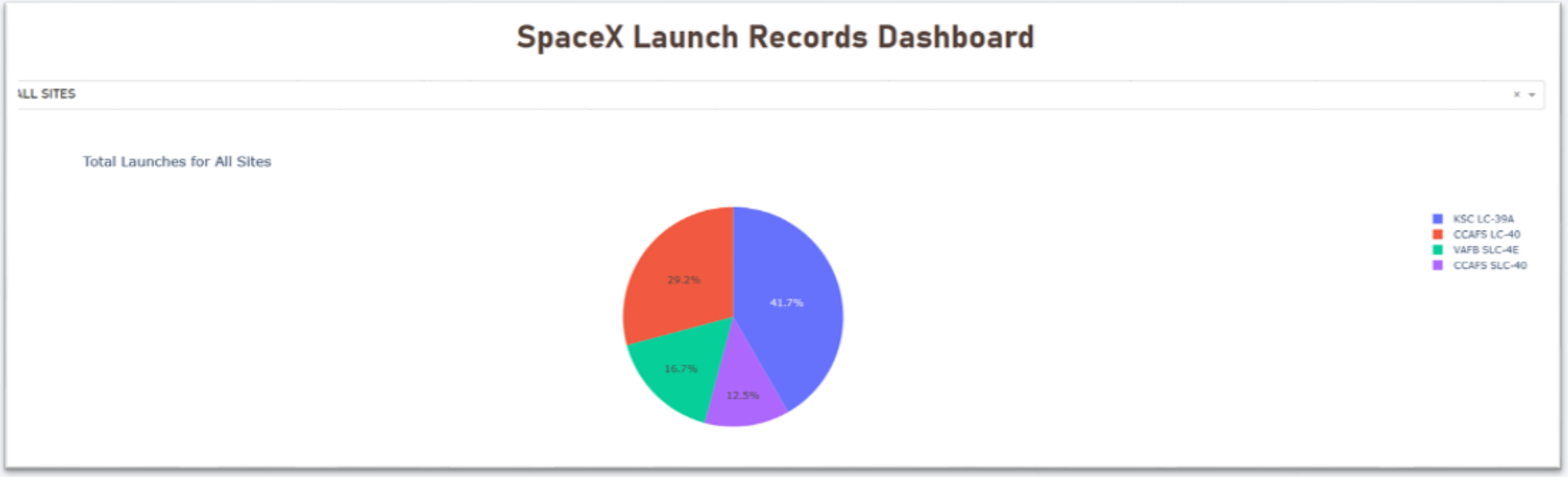
1. 0.58 km away from the closest highway
2. 0.86 km away from the closest coast
3. 1.27 km away from the closest railway
4. 51.5 km away from the closest big city




Section 4
Build a Dashboard with Plotly Dash

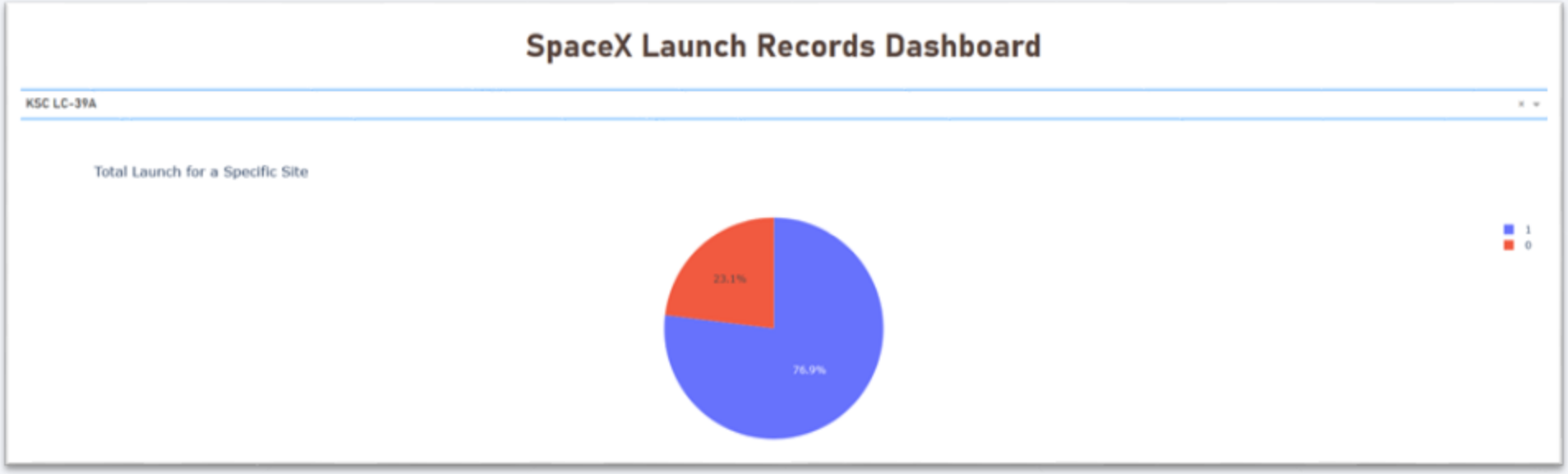
Success count for all sites

- KSC LC-39A has 10 successful missions, which is 41.7% of the total successful missions
- CCAFS LC-40 has 7 successful missions, which is 29.2% of the total successful missions



Success ratio for the most successful launch site

- The launch site with highest launch success ratio is the KSC LC-39A.
- The success ratio of KSC LC-39A is 76.9%.



Payload vs. Launch Outcome

- The success rate for all the launch sites and boosters for a payload range from 2500 to 9000 kg is represented on the right scatter point chart.

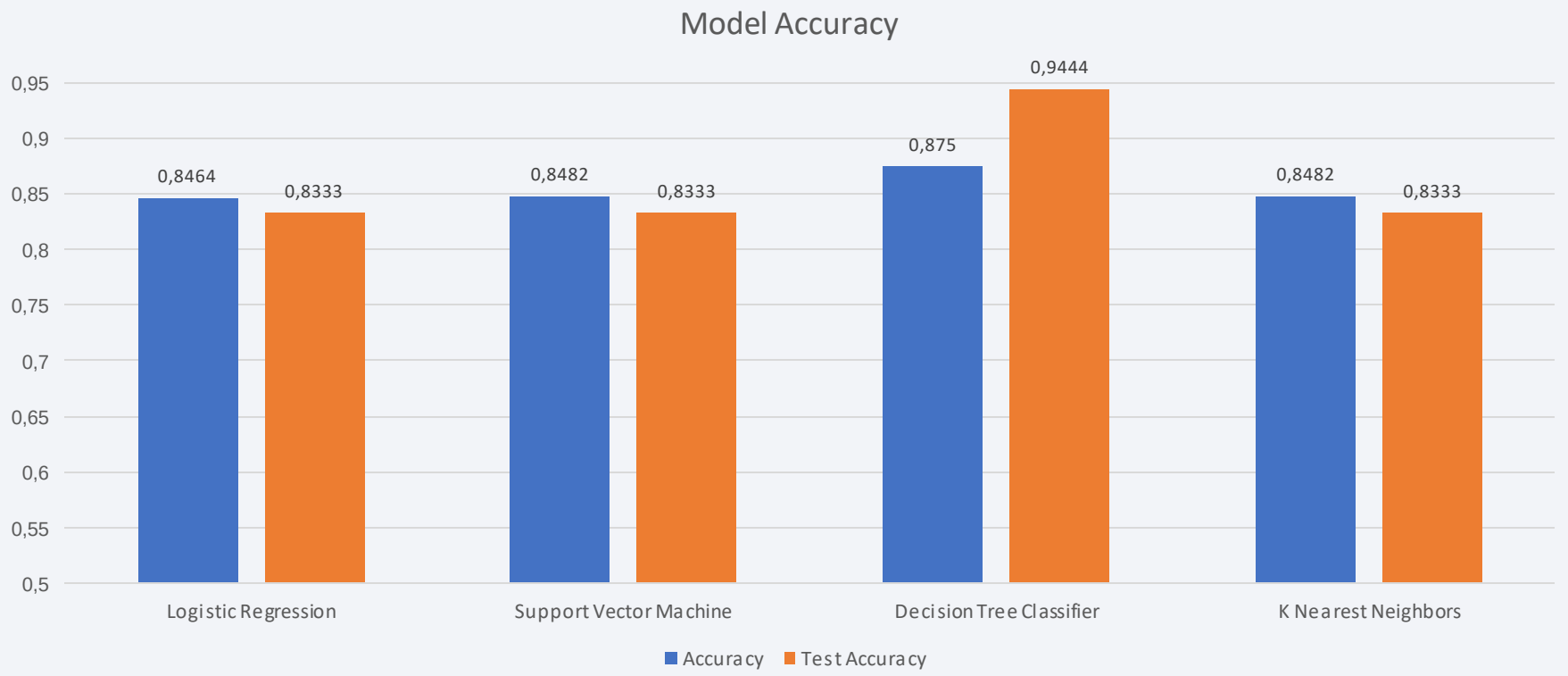
- For the full range of the payload, FT and B4 boosters seems to have the best success rates.

Section 5

Predictive Analysis (Classification)

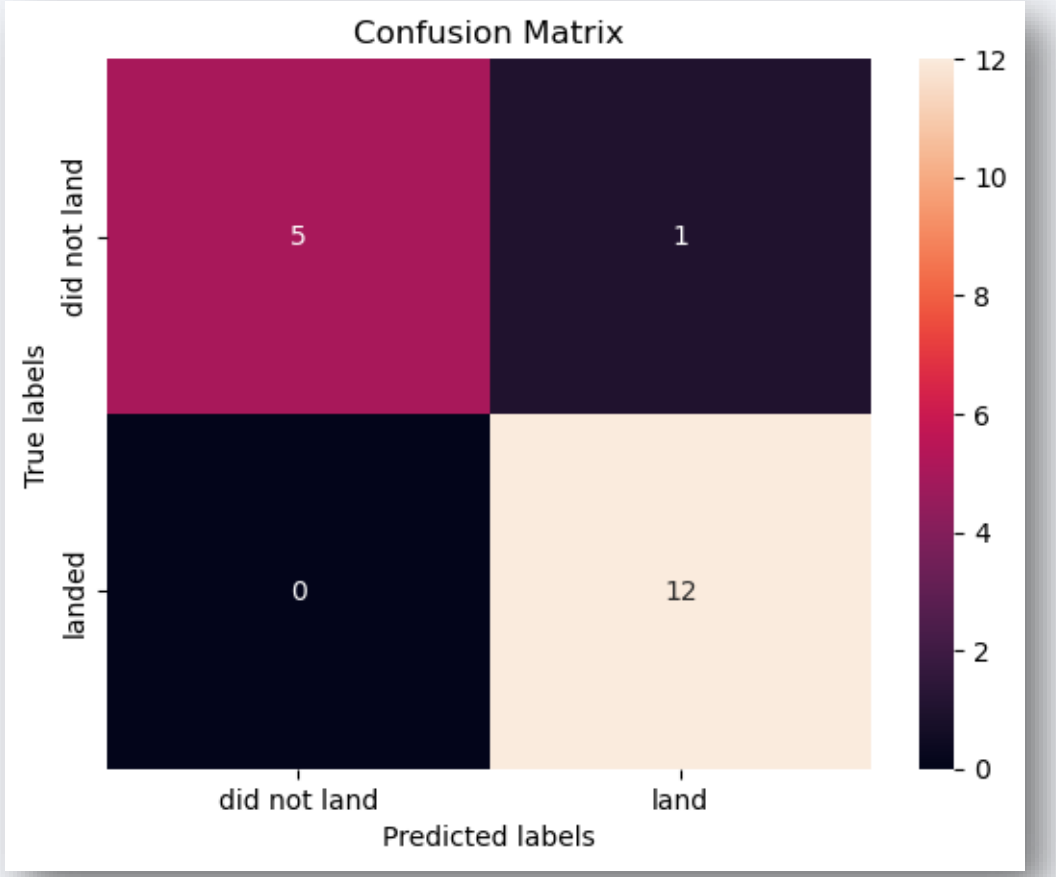
Classification Accuracy

- The best test accuracy performance was observed by the **Decision Tree Classifier**.



Confusion Matrix

- The confusion matrix for the optimal decision tree classifier shows that the model made only one false prediction out of 18 prediction on the test set. That is an amazing score but we have to consider that this might be a random event and we must not wait for such high scores while iterating the exact same model with new test sets.



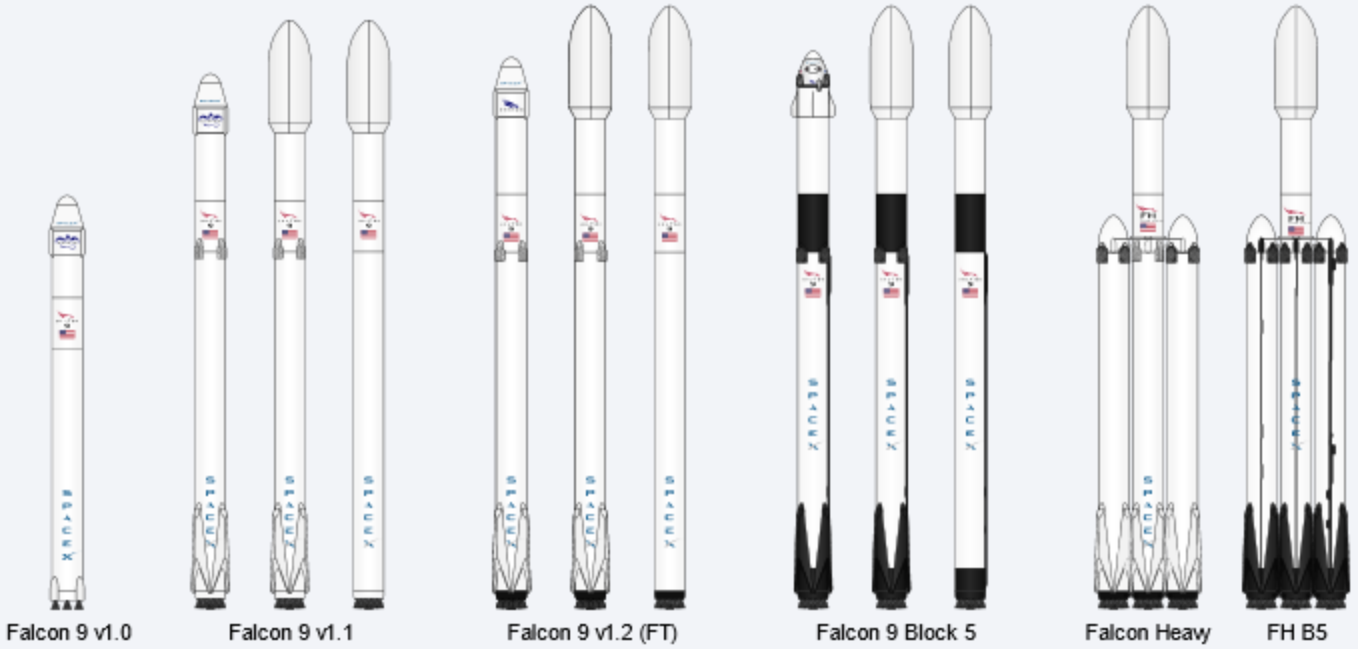
Conclusions

- Proximity to logistic infrastructures is a top priority.
- Launch sites might be near to the sea for security reasons.
- Orbit selection is an important task as it seems that there is a relation with the outcome of the missions.
- Every year the success rate gets better.
- KSC LC-39A is the most successful launch site with success ratio equal to 76.9%.
- The best classifier for this specific prediction task is the Decision Tree Classifier.

Appendix

- The whole project is available on:

https://github.com/Errikoskiladis/IBM-Applied-Data-Science-Capstone



Thank you!