IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jaana Jurvansuu
23.01.2022

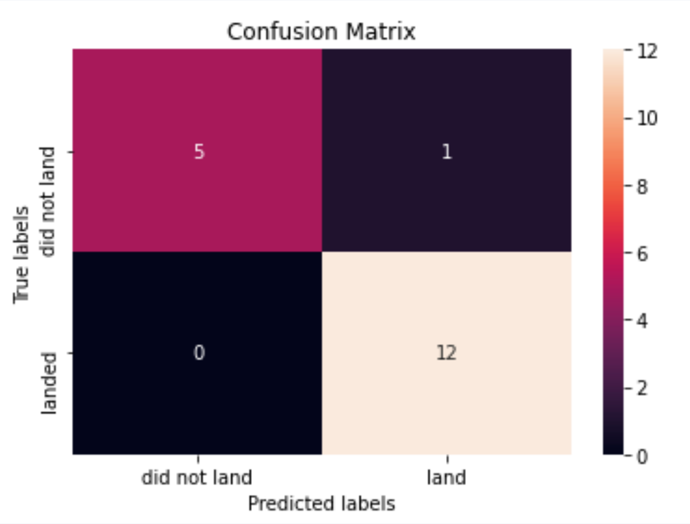
Outline

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

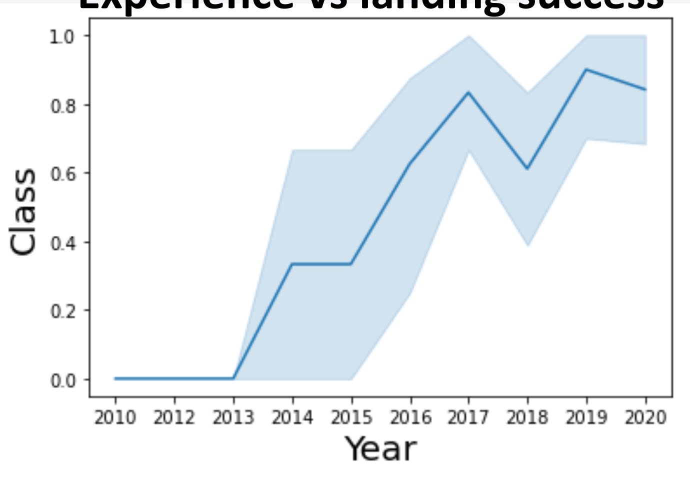
Executive Summary

- **How to recycle more first stage rocket parts?**
- **SpaceX launch data analysis.**
- Summary of methodologies:
    - Data sources: api.spacexdata.com/v4 and wikipedia.
    - Data wrangling, exploratory data analysis, feature engineering, interactive visual analytics, and machine learning predictions.
- Summary of all results:
    - Decision tree classifier is most accurate method to predict landing success from 17 different selected launch features.
    - Accuracy of decision tree classifier model is 0.94.
    - Launch experience correlates with landing success.

Decision tree classifier



Experience vs landing success



Introduction

Project background and context:

- SpaceY is a new commercial space flight company that would like to reduce the costs of space flights.
- Recycling first stage rocket parts, will reduce costs of rocket launches.

Problems to solve:

- Which launch and rocket features determine whether the first stage rocket landing is successful?
- Can these features be used to produce an accurate machine learning model to predict successful landings?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Data sources: api.spacexdata.com/v4 and en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches.
- Perform exploratory data analysis (EDA) using visualisation and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models:
    - Logistic regression
    - Vector machine object
    - K-nearest neighbours object
    - Decision tree classifier

Data Collection

- Data sets were collected by
  - API request to api.spacexdata.com/v4/
    - booster name, name of the launch site being used, the logitude and the latitude of the launch site, the mass of the payload, the rocket orbit, outcome of the landing, the type of the landing, number of flights with that core, whether gridfins were used, whether the core is reused, whether legs were used, the landing pad used, the block of the core, the number of times this specific core has been reused, and the serial of the core.
  - Scraping en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
    - Booster landing, launch outcome, flight No., dateTime, version Booster, launch site, payload, payload mass, orbit.

Data Collection – SpaceX API

The static response object with spaceX launch data: **“flight_number”** and **“date”**

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
```

API requested to api.spacexdata.com/v4 for **rockets, launchpads, payloads, and cores**

```
For example:
requests.get("https://api.spacexdata.com/v4/launchpads/"+str(x)).json()
```

Features of data_falcon9

```
FlightNumber
Date
BoosterVersion
PayloadMass
Orbit
LaunchSite
Outcome
Flights
GridFins
Reused
Legs
LandingPad
Block
ReusedCount
Serial
Longitude
Latitude
```

data_falcon9 .csv

The GitHub URL of the completed SpaceX API calls: github.com/jjurvans/SpaceY/blob/a2417f208ac82b2f30b5e6db2e33a57bf1832d7d/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Data was scraped from the wikipedia website and made into beautifulSoup object.

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
```

↓

Third table from the site was isolated (first_launch_table).

```
html_tables = soup.find_all('table')
first_launch_table = html_tables[2]
```

↓

The launch HTML table was parsed into dictionary, which was made into pandas data frame with features: Booster landing, launch outcome, flight No., dateTime, version Booster, launch site, payload, payload mass, orbit.

↓

spacex_web_scraped.csv

The GitHub URL of the completed web scraping notebook: github.com/jjurvans/SpaceY/blob/a2417f208ac82b2f30b5e6db2e33a57bf1832d7d/jupyter-labs-webscraping.ipynb

Data Wrangling

- Data with extra rocket boosters or multiple payloads were removed (data_falcon9.csv)
- Only Falcon 9 data was included (data_falcon9.csv)
- Empty cells were replaced with mean column values (data_falcon9.csv)
- Dates were replaced with datetime datatype (spacex_web_scraped.csv)
- Successful landing outcomes (True ASDS, True RTLS, and True Ocean) were labelled 1 and failed landing outcomes (None ASDS, False RTLS, None None, False ASDS, and False Ocean) were labelled 0.
- For predictive analysis values were OneHotEncoded by success class and standardised (feature engineering).

The GitHub URL of the completed data wrangling notebook: https://github.com/jjurvans/SpaceY/blob/7f531adec8b5aa34094907620d8f0b2d909c686b/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualisation

The success of landing was studied by year and orbit and with relationships between:

- **FlightNumber vs. PayloadMass and Orbit**
  - Does the success of landings depend on experience, payload or orbit?
- **LaunchSite vs. FlightNumber and PayloadMass**
  - Does the success of landings depend on launch site?
- **PayloadMass vs. Orbit**
  - Does the success of landings depend on orbit

The GitHub URL of the completed EDA with data visualisation notebook: https://github.com/jjurvans/SpaceY/blob/7f531adec8b5aa34094907620d8f0b2d909c686b/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- The names of unique launch sites.
- 5 records where launch sites begin with the string 'CCA'.
- The total payload mass carried by boosters launched by NASA.
- The average payload mass carried by booster version F9 v1.1.
- The date when the first successful landing outcome in ground pad was acheived.
- List 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.
- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- The 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.

[The GitHub URL]() of the completed EDA with SQL notebook: https://github.com/jjurvans/SpaceY/blob/7f531adec8b5aa34094907620d8f0b2d909c686b/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

- Each launch site was marked by folium.Circle and folium.Marker to the map.
- MarkerCluster object was added to each launch sites to show amount of successful and unsuccessful landings.
- MousePosition indicator was added to the map and used to calculate distances between launch sites and closest coastline, city, railway and highway. Some distances and distance lines were marked into the map.

The GitHub URL of the completed Folium notebook: https://github.com/jjurvans/SpaceY/blob/7f531adec8b5aa34094907620d8f0b2d909c686b/lab_jupyter_launch_site_location.ipynb

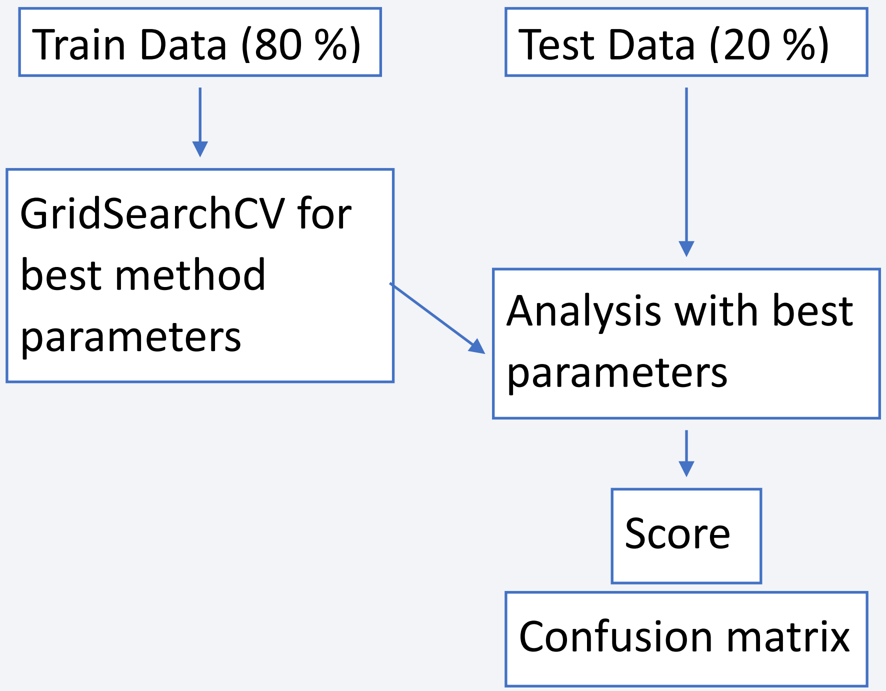
Build a Dashboard with Plotly Dash

- SpaceX launch records dashboard has two parts that took feed from dropdown selection to choose all launch sites or individual sites.
  - The Pie chart shows the total launch success rate for each launch sites or ratio of successful and unsuccessful launches for each site.
  - The scatter plot shows correlation between Payload and landing success for all or individual sites. The payload range can be adjusted by a slider.

The GitHub URL of the completed Folium notebook: https://github.com/jjurvans/SpaceY/blob/7f531adec8b5aa34094907620d8f0b2d909c686b/spaceY_dash_app.py

Predictive Analysis (Classification)

- The features explaining the landing success were from data_falcon9.
- Training set had 72 cases (80 % of the data) and test set had 18 cases (20 % of the data).
- Best hyperparameters were found for Logistic regression, Vector machine object K-nearest neighbours object, and Decision tree classifier analyses. The methods' performance was evaluated by test data mean accuracy (score) and confusion matrix.



The GitHub URL of the completed Folium notebook: https://github.com/jjurvans/SpaceY/blob/e17eaa446d10fe3422a19b5448b0a82f7d597556/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

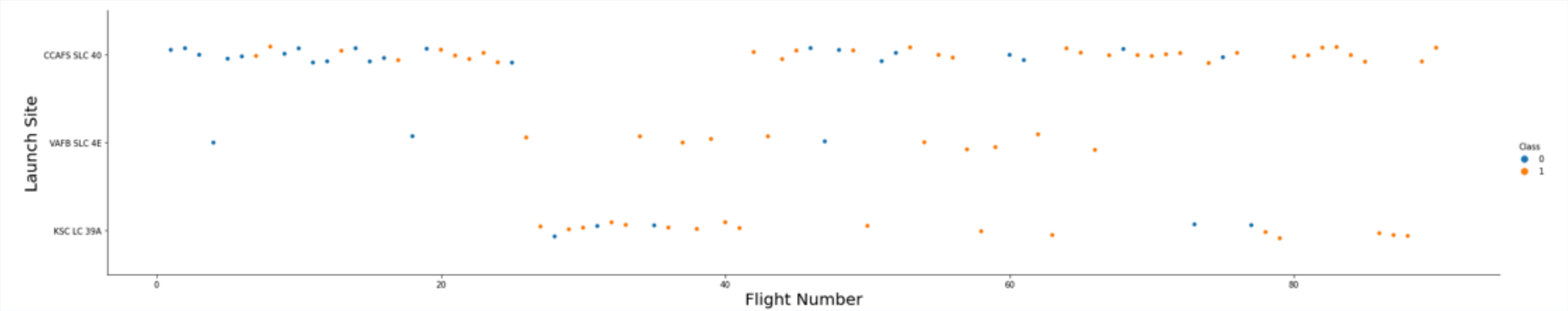


Figure 1. A scatter plot of Flight Number vs. Launch Site. The launch sites are CCAFS LC-40, KSC LC-39A and VAFB SLC 4E. Class 0 in blue denotes unsuccessful landings and class 1 in orange successful landings. The plot show that when the flight number increases the success of landings becomes more frequent.

Payload vs. Launch Site

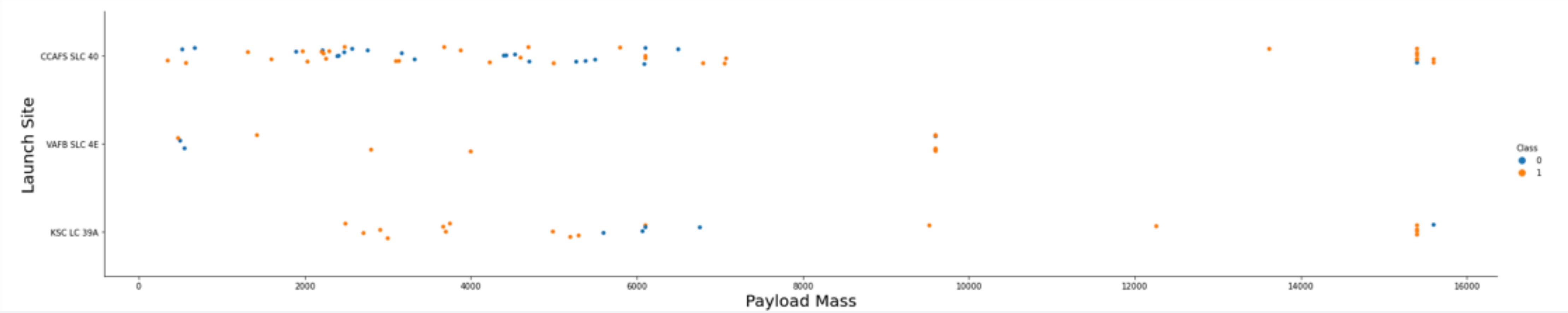
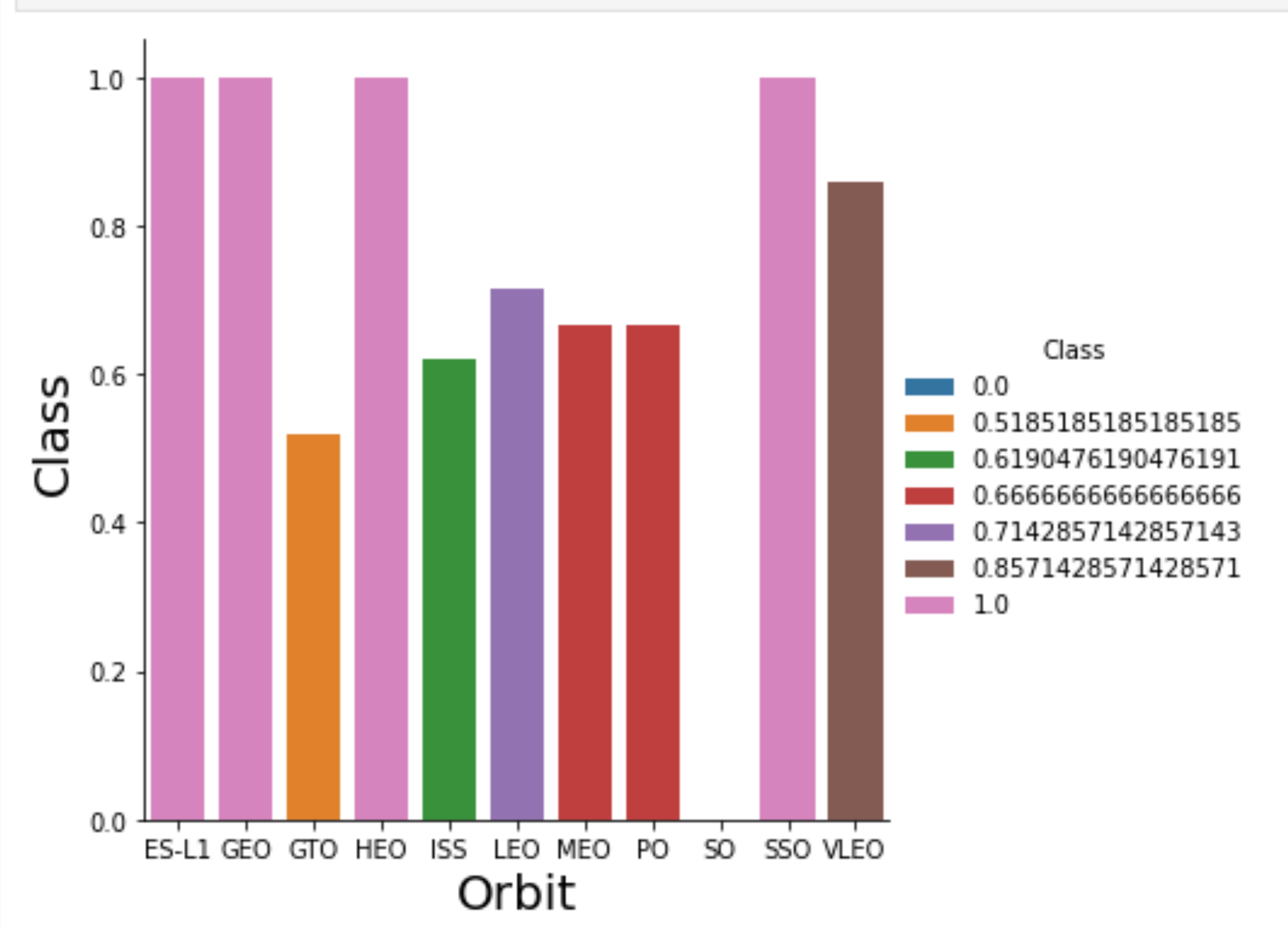


Figure 2. A scatter plot of Payload vs. Launch Site. The launch sites are CCAFS LC-40, KSC LC-39A and VAFB SLC 4E. Class 0 in blue denotes unsuccessful landings and class 1 in orange successful landings. The plot shows that there is no clear pattern between weight of payload and success of landing. The results might be bias by higher amount of light payload flights.

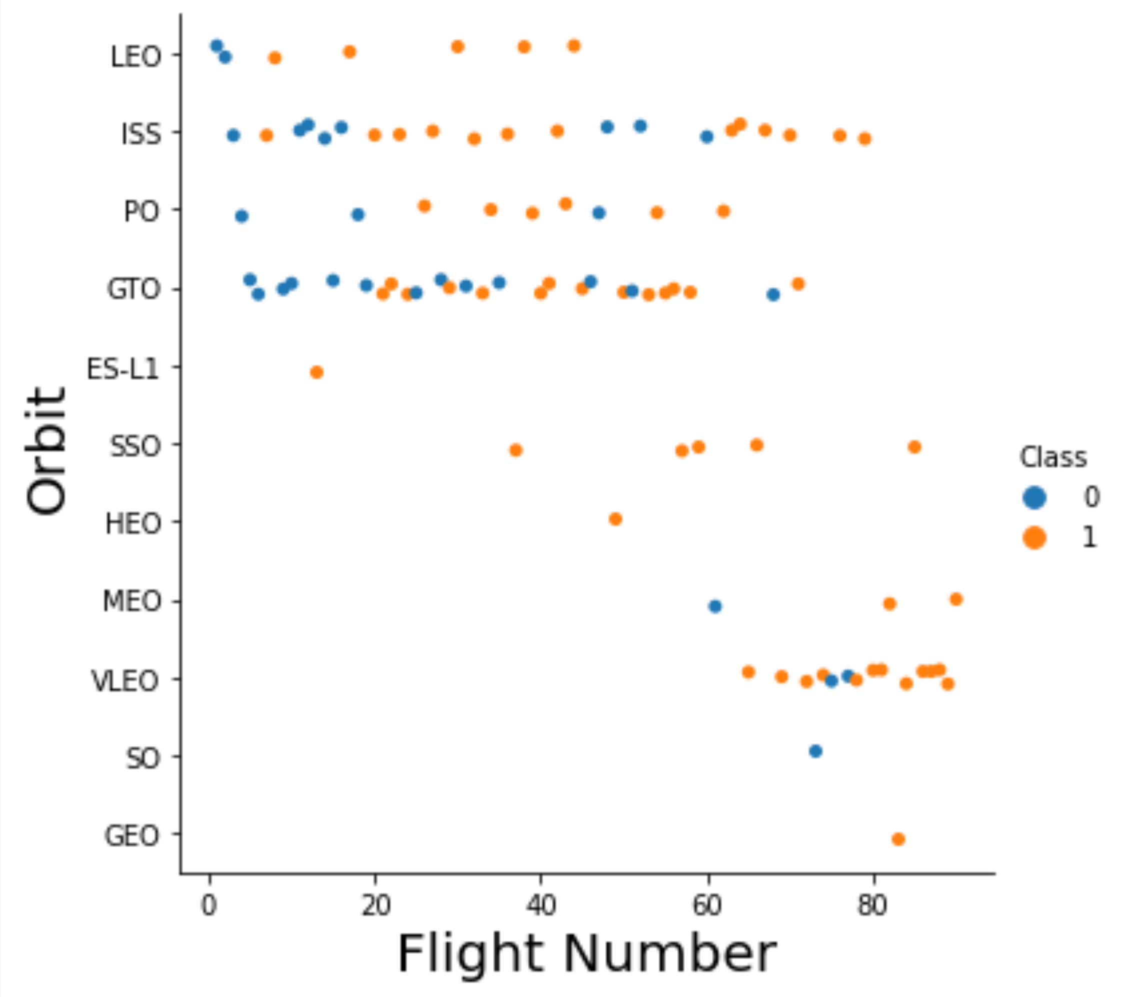
Success Rate vs. Orbit Type

Figure 3. A bar chart for the success rate of each orbit type. Class denotes the average rate of landing success, i.e., class 1 means all the landings have been successful. Landings from orbits ES-L1, GEO, HEO, and SSO were all successful, although GEO, ES-L1 and HEO had only one flight. GTO, ISS and VLEO had most flights (over 14) and thus most reliable data. From these VLEO has the highest success rate (0.86) and it has the most closest orbit.



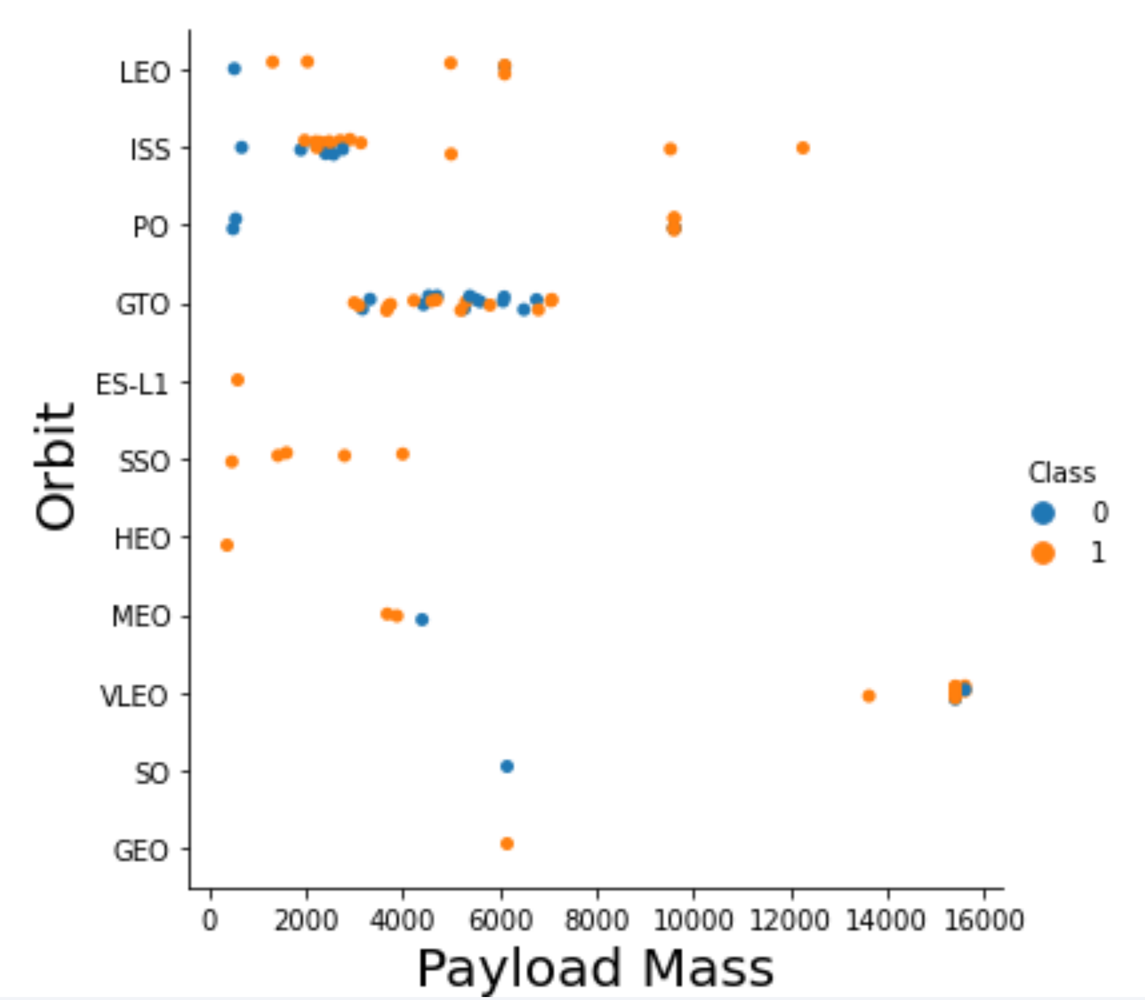
Flight Number vs. Orbit Type

Figure 4. A scatter plot of Flight number vs. Orbit type. Class 0 in blue denotes unsuccessful landings and class 1 in orange successful landings. The plot shows that VLEO orbit is used only in later flights, which may explain their success rate (see figure 3). LEO, ISS, PO, and GTO orbit flight have been used from early on, which may explain their lower success rate.



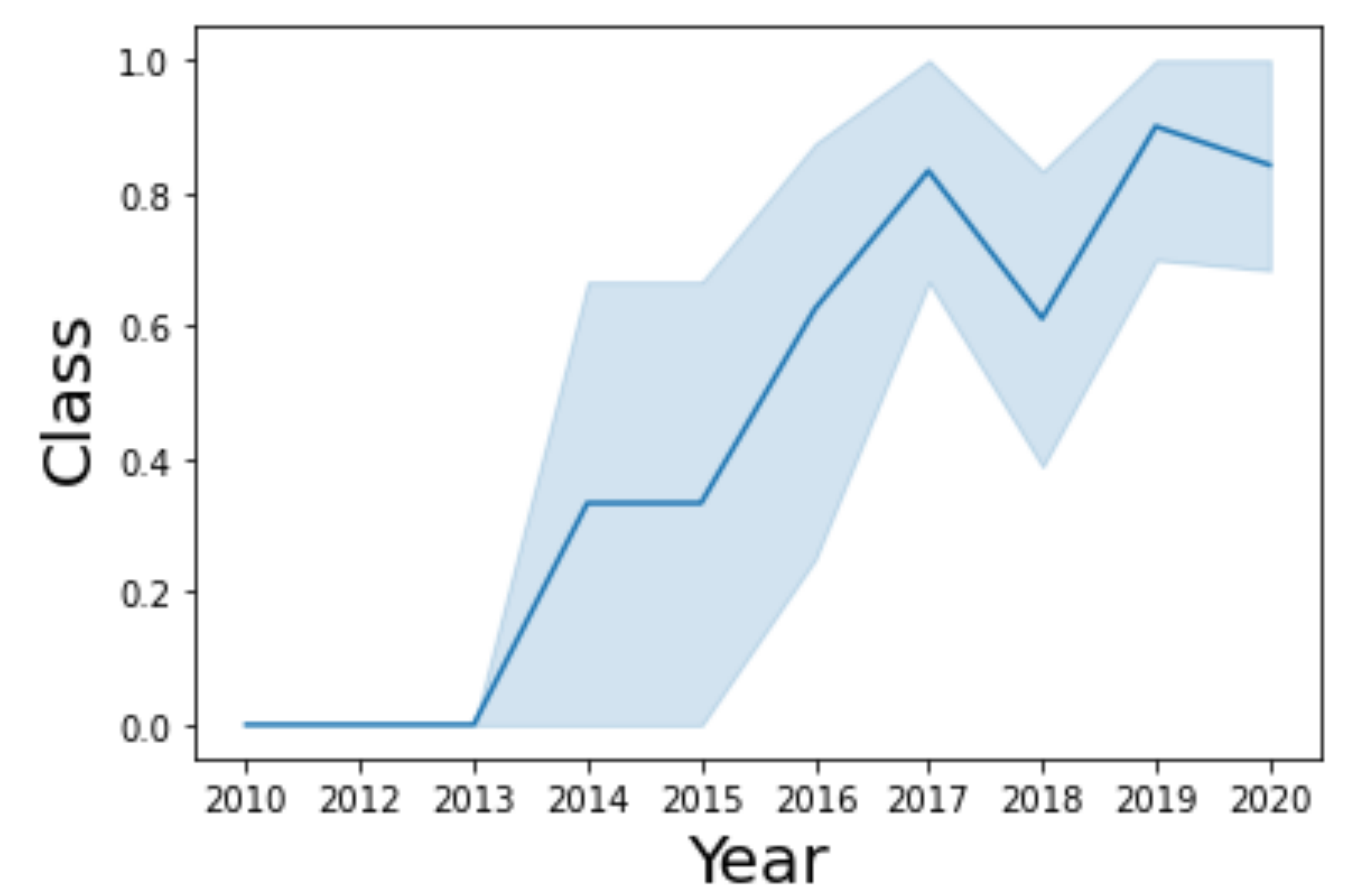
Payload vs. Orbit Type

Figure 5. A scatter plot of payload vs. orbit type. Class 0 in blue denotes unsuccessful landings and class 1 in orange successful landings. The plot shows that there is only a few flights with payload mass higher than 10000 kg in any orbit. However VLEO flight are all over 12000 kg. With exception of GTO orbit, it seems that the higher payload reads into successful landing.



Launch Success Yearly Trend

Figure 6. A line chart of yearly average success rate of the landings (Class). The dark blue line shows the average success rate and the light blue the variation. The chart shows that landing success tends to increase with more experience.



All Launch Site Names

- The unique launch sites are:
- CCAFS SLC-40 and LC-40 are Cape Canaveral Air Force Station Space Launch Complex 40 in Florida, US.
- KSC LC-39 is Kennedy Space Center Launch Complex 39 in Florida, US.
- VAFB SLC-4E is Vandenberg Air Force Base Space Launch Complex 4 East in California, US.

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites 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 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 total payload carried by boosters from NASA is 45 595 kg.

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(payload_mass__kg_) from spacextbl where customer='NASA (CRS)';
```

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928 kg.

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

```
%sql select avg(payload_mass__kg_) from spacextbl where booster_version='F9 v1.1';
```

First Successful Ground Landing Date

- The first successful landing outcome on ground pad was on Tuesday 22.12.2015.

```
%sql select min(date) from spacextbl where landing__outcome like 'Success%';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- List of names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000:

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

Total Number of Successful and Failure Mission Outcomes

- The total number of successful is and failure mission outcomes is 99.

```
%sql select count(mission_outcome) from spacextbl group by mission_outcome;
```

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass of 15600 kg:

booster_version
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:

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

- Rank of 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:

DATE	landing__outcome	RANK
2013-09-29	Uncontrolled (ocean)	1
2014-09-21	Uncontrolled (ocean)	1
2015-12-22	Success (ground pad)	3
2016-07-18	Success (ground pad)	3
2017-02-19	Success (ground pad)	3
2016-04-08	Success (drone ship)	6
2016-05-06	Success (drone ship)	6
2016-05-27	Success (drone ship)	6
2016-08-14	Success (drone ship)	6
2017-01-14	Success (drone ship)	6
2015-06-28	Precluded (drone ship)	11
2012-05-22	No attempt	12
2012-10-08	No attempt	12
2013-03-01	No attempt	12
2013-12-03	No attempt	12
2014-01-06	No attempt	12
2014-08-05	No attempt	12
2014-09-07	No attempt	12
2015-03-02	No attempt	12
2015-04-27	No attempt	12
2017-03-16	No attempt	12
2010-06-04	Failure (parachute)	22
2010-12-08	Failure (parachute)	22
2015-01-10	Failure (drone ship)	24
2015-04-14	Failure (drone ship)	24
2016-01-17	Failure (drone ship)	24
2016-03-04	Failure (drone ship)	24
2016-06-15	Failure (drone ship)	24
2014-04-18	Controlled (ocean)	29
2014-07-14	Controlled (ocean)	29
2015-02-11	Controlled (ocean)	29

Section 4

Launch Sites Proximities Analysis

Locations of the SpaceX launch sites

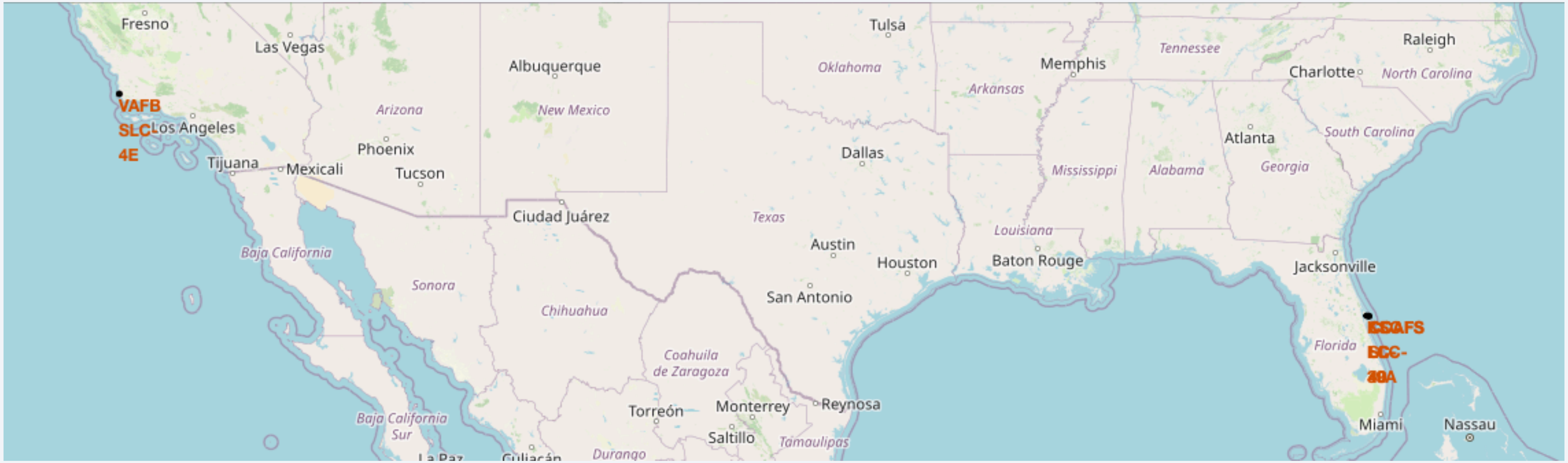


Figure 7. Locations of the SpaceX launch sites. Launch sites are marked in the map in red circle and name of the site. The launch sites are not on equatorial line but they are close proximity to the coast to enable landing to sea.

Landing outcomes on the map

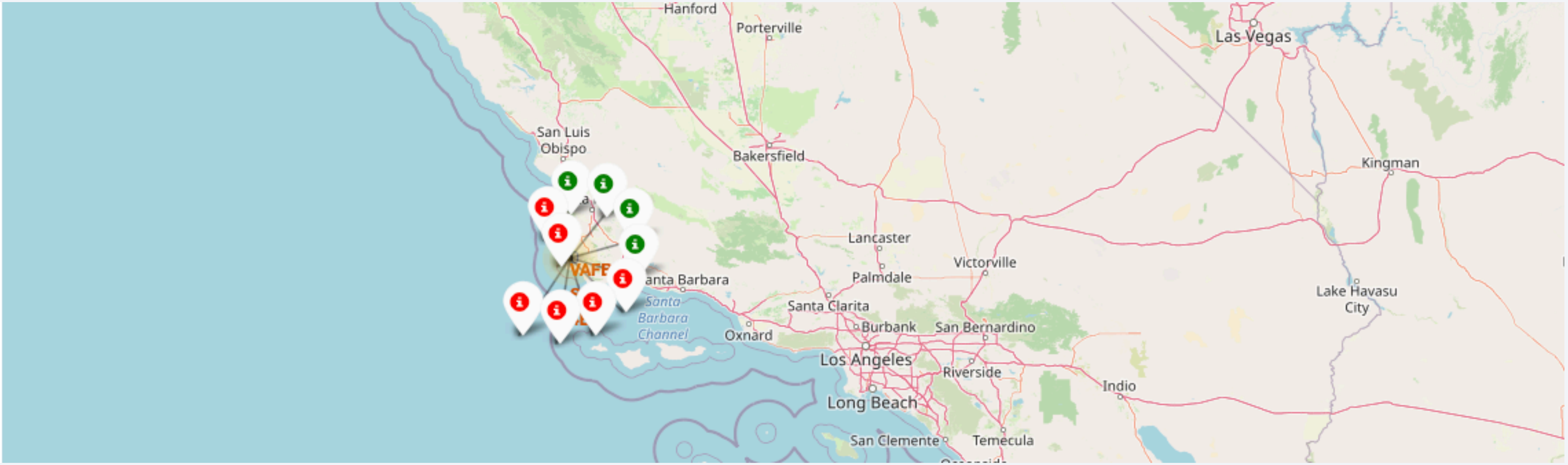


Figure 8. Landing outcomes of the SpaceX launch sites at VAFB SLC-4E. The successful landings are shown in green info circle and the failed landings are shown in red info circles.

Launch site distances to major cities and highways

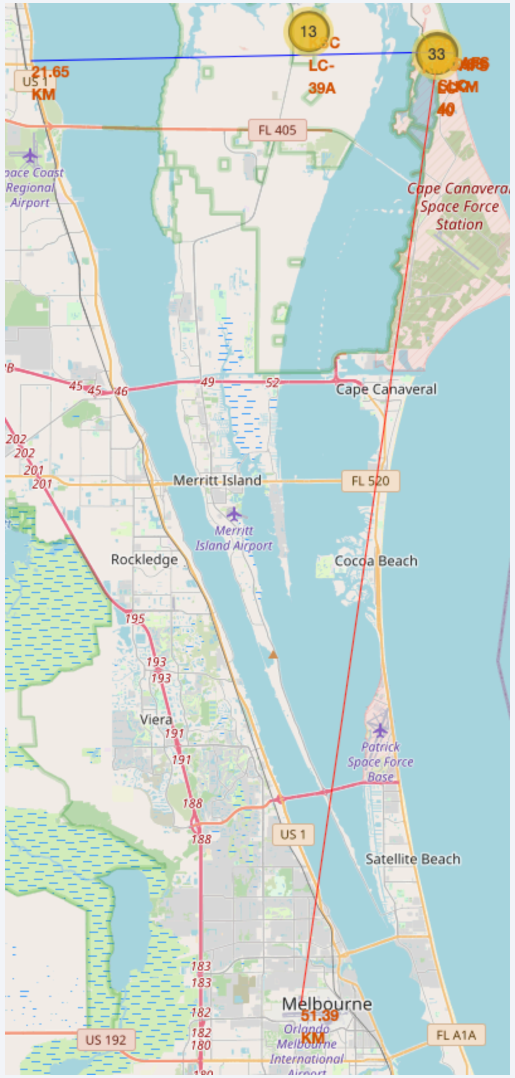
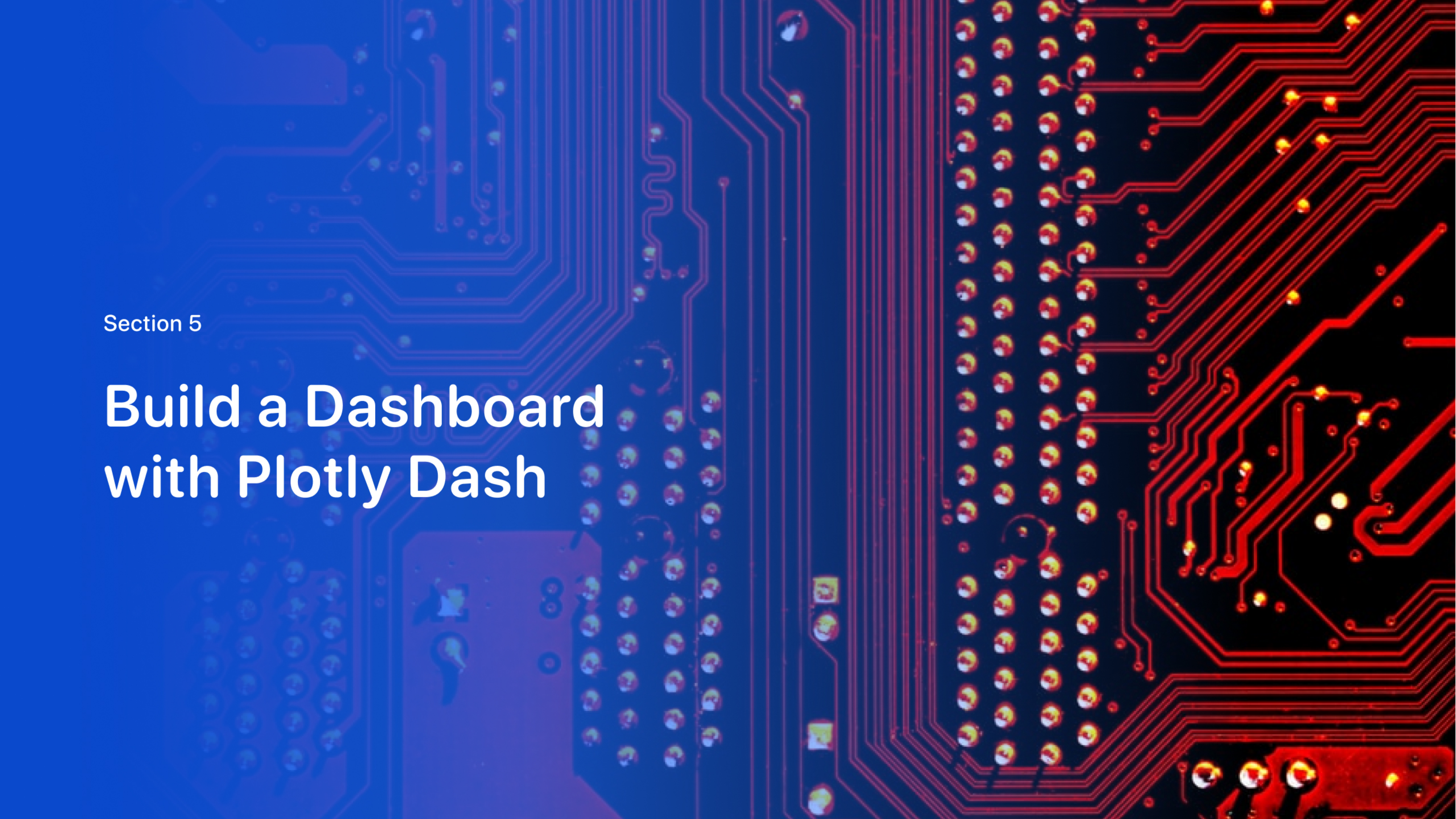


Figure 9. SpaceX launch sites CCAFS SLC-40 distance to the nearest city and highway. The distance from the CCAFS SLC-40 launch site to Melbourne, Florida, US, is shown in red line and the distance to the city is 51.4 km. The distance from the launch site to Washington Avenue is shown in blue line and the distance to the highway is 21.7 km.



Section 5

Build a Dashboard with Plotly Dash

Example of the Dashboard: all the launch sites

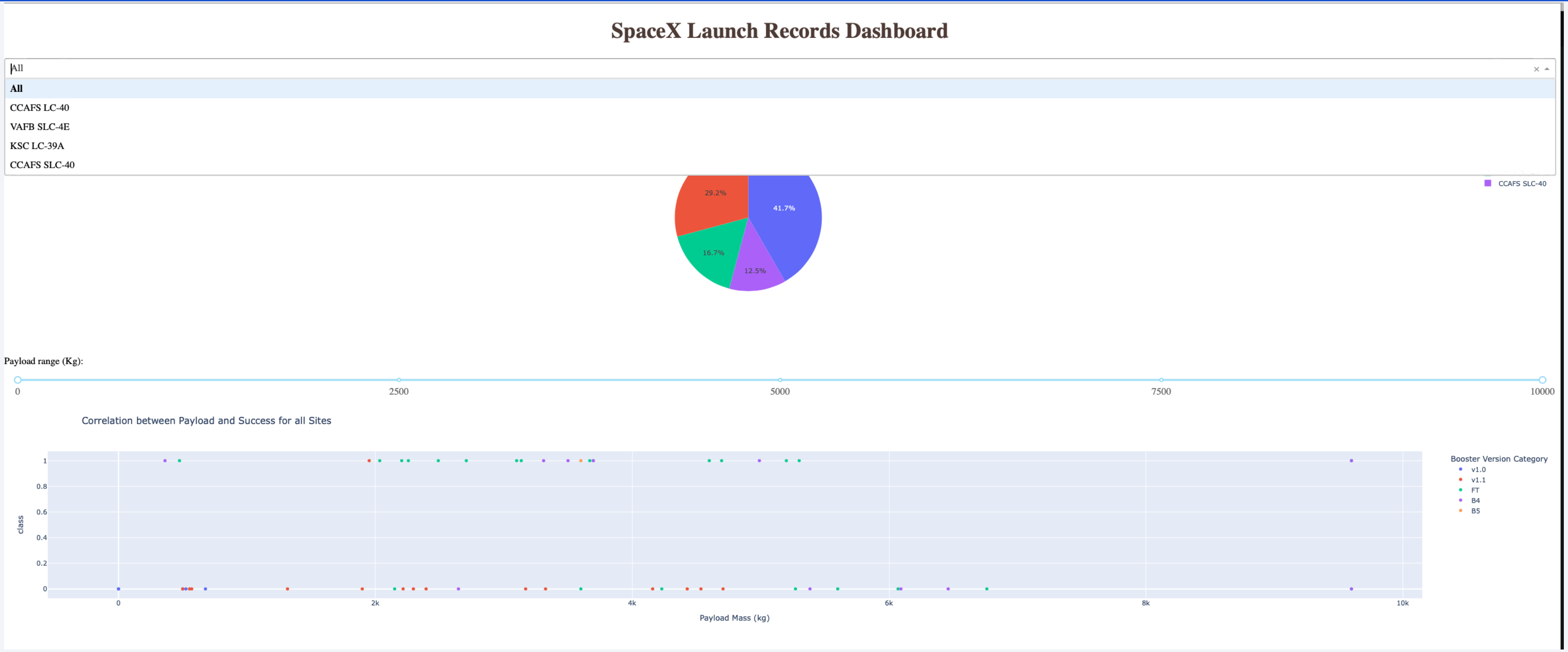


Figure 10. SpaceX launch records dashboard with all launch sites selected. The pie chart shows total success of landings for all the sites. The scatter plot shows the correlation between payload and landing success for all sites and all payloads by booster version.

Is the launch site with highest launch success ratio

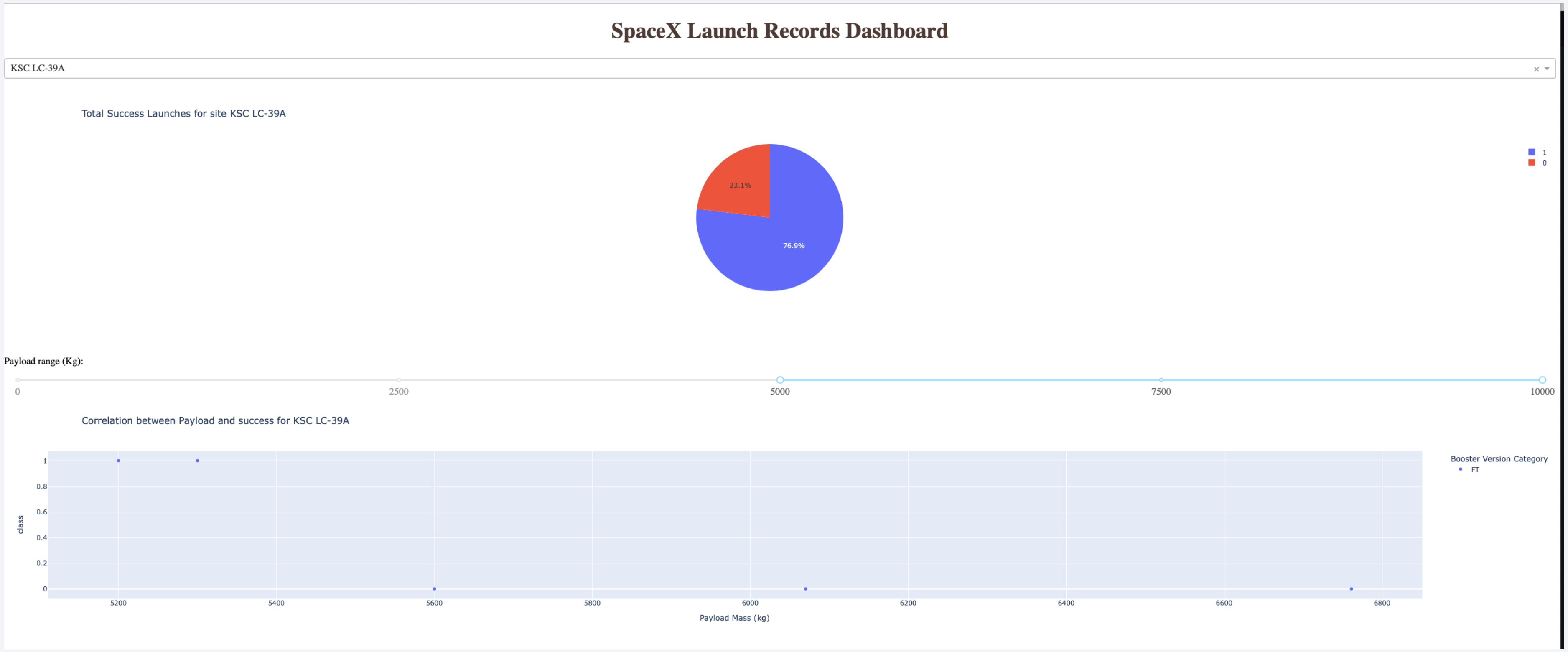


Figure 10. SpaceX launch records dashboard for KSC LC-39A launch site. The pie chart shows percentage of successful of landings in blue and failed landings in red. The scatter plot shows the correlation between payload and landing success with payload range 5000-10000 kg.

<Dashboard Screenshot 3>

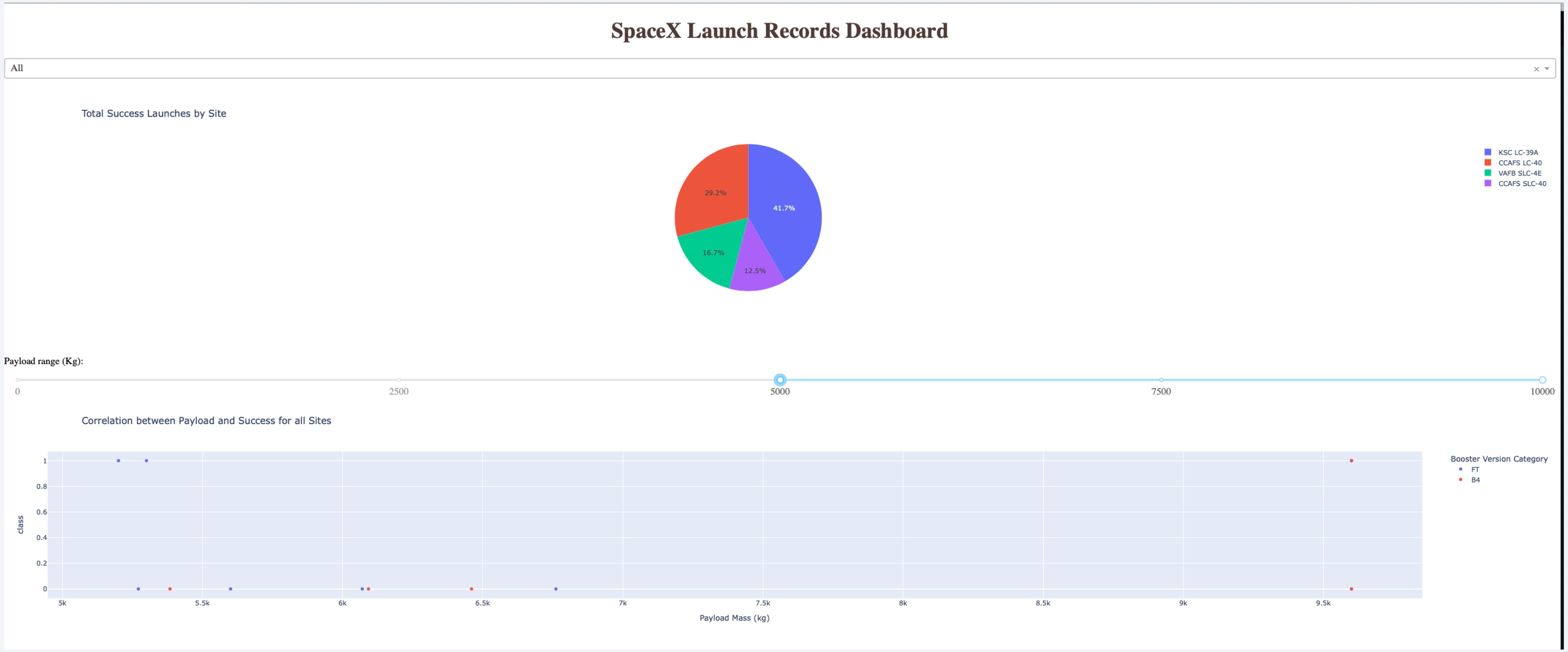


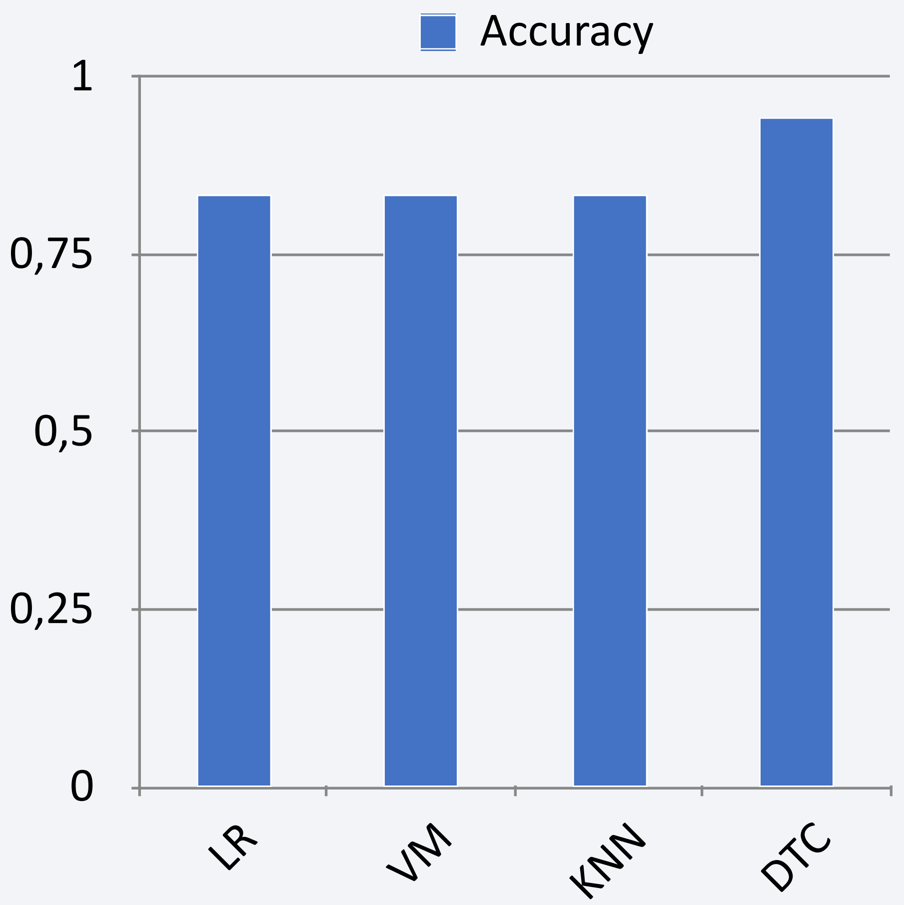
Figure 11. SpaceX launch records dashboard with all launch sites selected by payload range of 5000-10000 kg. The pie chart shows total success of landings for all the sites. The scatter plot shows the correlation between payload and landing success for all sites by booster version for payload range of 5000-10000 kg.

Section 6

Predictive Analysis (Classification)

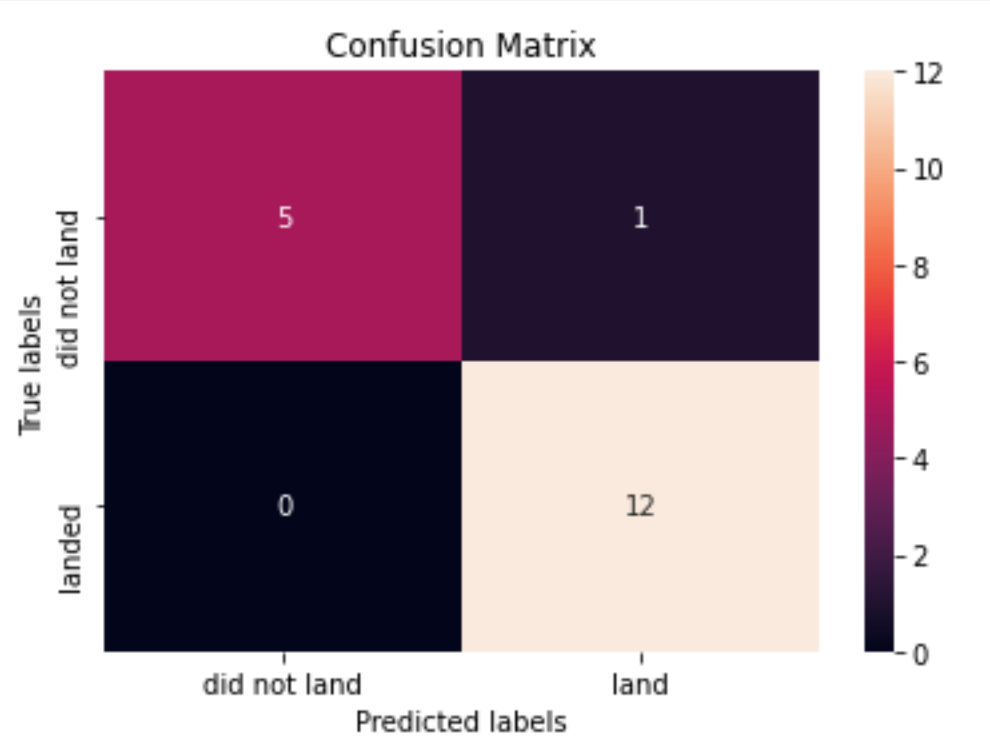
Classification Accuracy

Figure 12. Classification accuracy of the methods. Logistic regression (LC), Vector machine object (VM), and K-nearest neighbours object (KNN) were all equally accurate (0,83). Decision tree classifier was slightly more accurate (0,94) than the three other models to classify the test data.



Confusion Matrix

Figure 13. Confusion matrix of the most accurate model: the Decision tree classifier. The methods predicts wrongly only one landing for successful landing that actually failed landing.



Conclusions

- **The Decision tree classifier is the most accurate model to predict landing success from the data available.**
- The test set was very small (18 cases) and more testing would be useful to further validate the model.
- Single most important feature to predict landing success seems to be launch experience (the feature: FlightNumber).
- More features could make the model more accurate.

Appendix

- github.com/jjurvans/SpaceY.git

Thank you!