Winning Space Race with Data Science

TAO ZHE (EVAN) WU
2023-02-26

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection using web scraping and Space X API;
  - Exploratory Data Analysis (EDA), including data wrangling, data visualization and interactive visual analytics;
  - Machine Learning Prediction.
- Summary of all results
  - It was possible to collect valuable data from public sources;
  - EDA allowed to identify which features were the best to predict the success of launchingsl;
  - Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way, using all collected data.

Introduction

- Project background and context
  - The objective is to evaluate the viability of the new company Space Y to complete with Space X.
- Problems you want to find answers
  - The best way to estimate the total cost for launches, by predicting successful landings of the first stage of rockets;
  - The best place to make launches.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data from Space X was obtained from two sources:
    - 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

Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- 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 suing different combinations of parameters.

Data Collection

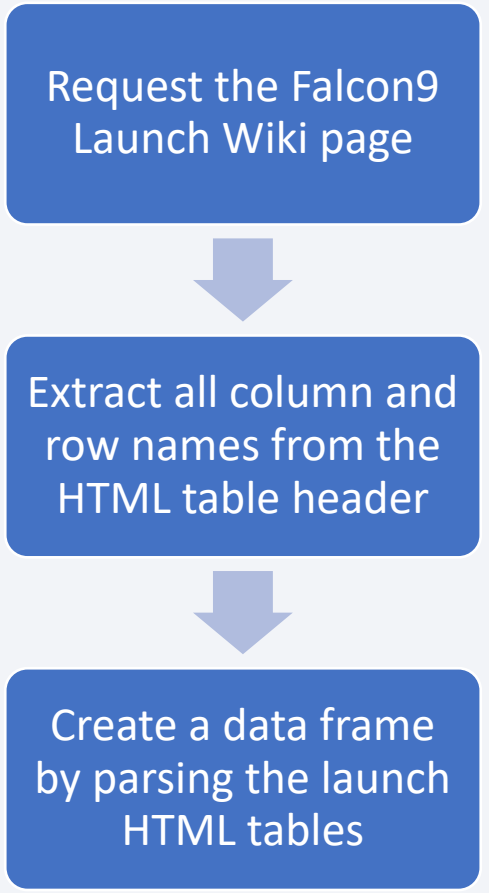
- Data sets were collected:
  - From Space X API (https://api.spacexdata.com/v4/rockets/)
  - From Wikipedia (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches) by web scraping.

Data Collection – SpaceX API

- SpaceX offers a public API from where data can be obtained and used;
- This API was used according to the flow chart beside and then data is persisted.
- Source code: https://github.com/wttz1212/Applied-Data-Science-Capstone

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data were downloaded from Wikipedia according to the flowchart and then persisted.
- Source code: https://github.com/wttz1212/Applied-Data-Science-Capstone/blob/main/jupyter_labs_webscraping.ipynb



Data Wrangling

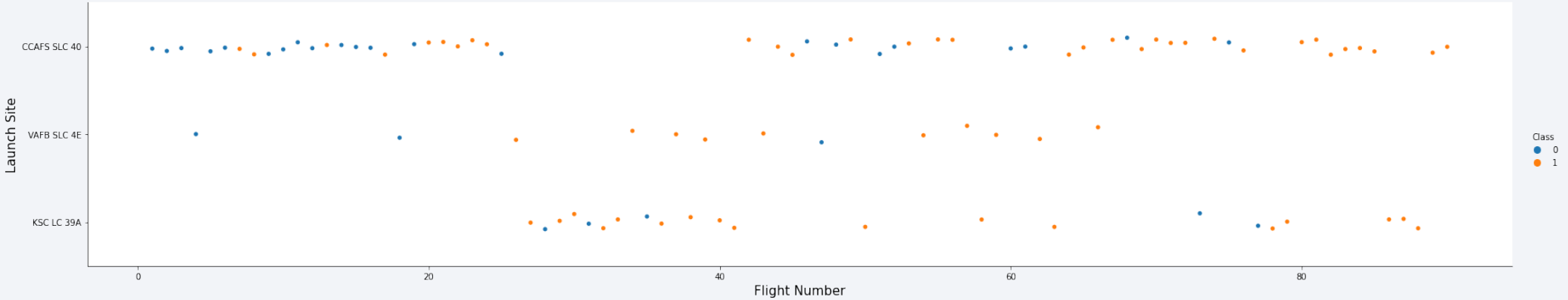
- Initially some Exploratory Data Analysis (EDA) was performed on the dataset;
- Then the launches per site, occurrences of each orbit and occurrences of mission outcome per orbit were calculated.
- Finally, the landing outcome label was created from Outcome column.

- Source code:

https://github.com/wttz1212/Applied-Data-Science-Capstone/blob/main/labs_jupyter_spacex_Data_wrangling.ipynb

EDA with Data Visualization

- To explore data, scatterplots and barpolts were used to visualize the relationship between paris of features:
  - Payload Mass vs Flight Number, Launch Site vs Flight Number, Launch Site vs Payload Mass, Orbit vs Flight Number, Payload vs Orbit



- Source code: https://github.com/wttz1212/Applied-Data-Science-Capstone/blob/main/jupyter_labs_eda_dataviz.ipynb

EDA with SQL

- The following SQL queriers were performed:
  - Names of the unique launch sites in the space mission;
  - Top 5 launch sites whose name begin with the string 'CCA';
  - Total payload 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 6000 kg;
  - Total number of the boosters which have carried the maximum payload mass;
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for in 2015 year;
  - Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-05 and 2017-03-20 in descending order.
- Source code: https://github.com/wttz1212/Applied-Data-Science-Capstone/blob/main/jupyter_labs_eda_sql_coursera.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate launch sites;
  - Circles indicate highlighted areas around specific coordinates like NASA Johnson Space Center;
  - Marker clusters indicate groups of event in each coordinate like launches in a launch site;
  - Lines are used to indicate distances between two coordinates.

- Source code: https://github.com/wttz1212/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- The following plots/graphs and interactions were added to a dashboard:
  - Percentage of launches by site
  - Payload range

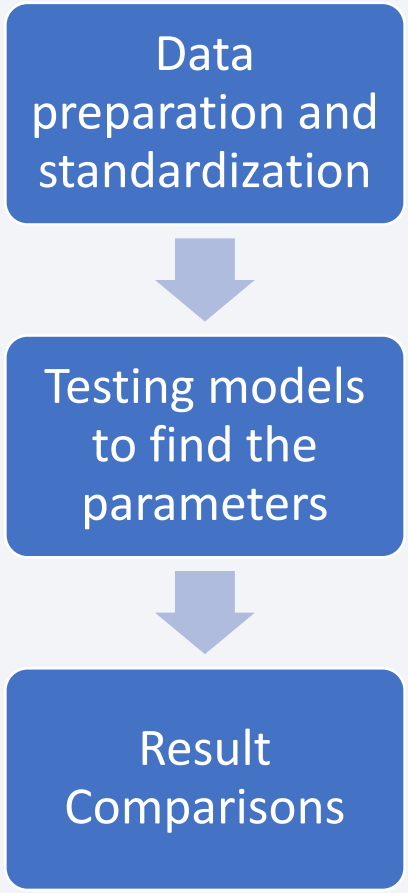
- Those plots and interactions allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is the best place to launch according to payloads.

- Source code: https://github.com/wttz1212/Applied-Data-Science-Capstone/blob/main/dash_interactivity.py

Predictive Analysis (Classification)

- Four classification models (Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors) were compared to find the best performing classification model.

- Source code: https://github.com/wttz1212/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.ipynb

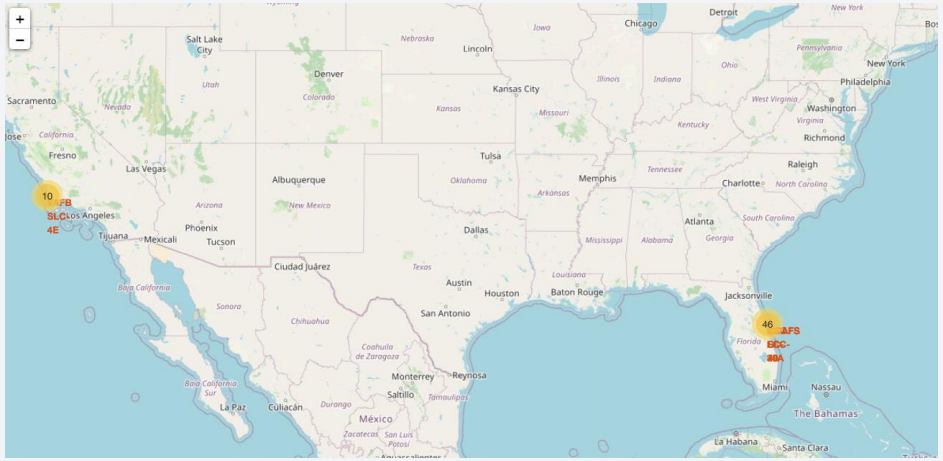


Results

- Exploratory data analysis results:
  - Space X uses 4 different launch sites;
  - The first launches were done by Space X itself and NASA;
  - The average payloads of F9 v1.1 booster is 2928 kg;
  - The first success landing outcome happened in 2015 fiver year 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 in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
  - The number of landing outcomes became as better as years passed.

Results

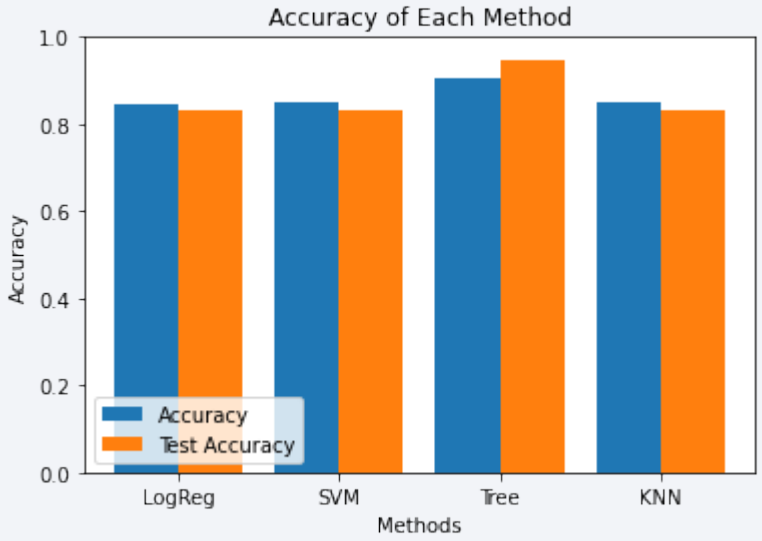
- Using interactive analysis, it was possible to identify that launch sites are in safe places near sea and have a good logistic infrastructure around.
- Most launches happens at east coast launch sites.



Results

- Predictive Analysis showed that Decision Tree Classifier has the highest accuracy (90.36%) and accuracy for test (94.44%) to predict successful landings.

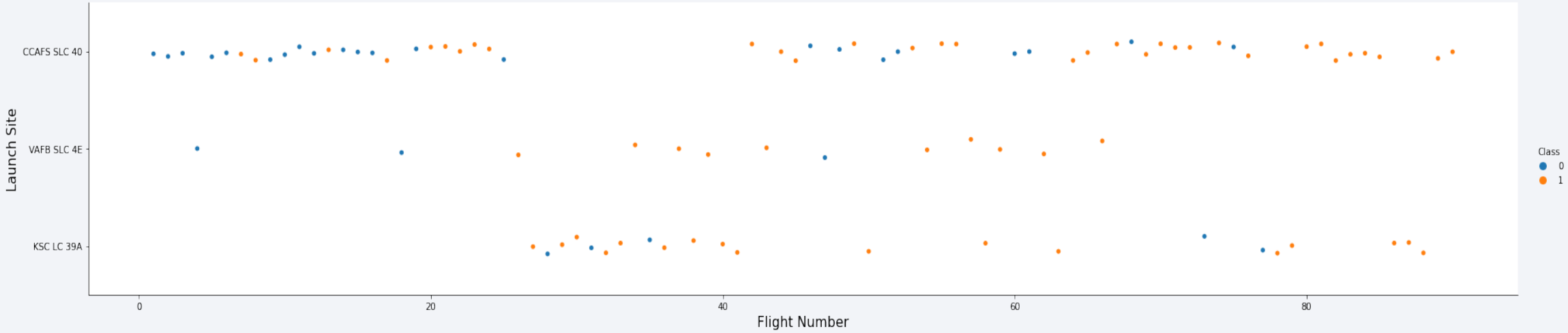
	Model	Accuracy	Test Accurracy
0	LogReg	0.8464	0.8333
1	SVM	0.8482	0.8333
2	Tree	0.9036	0.9444
3	KNN	0.8482	0.8333



Section 2

Insights drawn from EDA

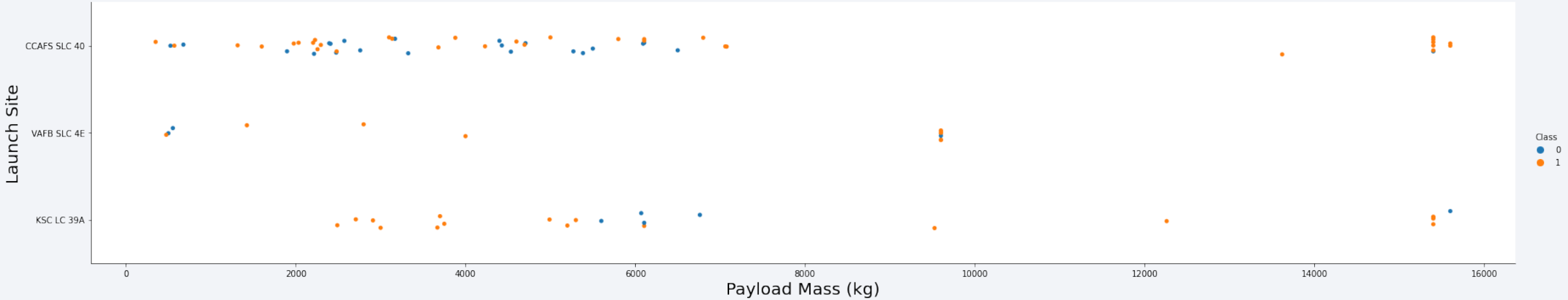
Flight Number vs. Launch Site



- Total number of flights for CCAFS SLC 40 are 55:
  - 33 successful / 22 failed
- Total number of flights for VAFB SLC 4E are 13
  - 10 successful / 3 failed
- Total number of flights for KSL LC 39A are 22:
  - 17 successful / 5 failed

- Preliminary analysis shows that launch sites KSL LC 39 A and VAFB SLC 4E are more successful than CCAFS SLC (having more than 76% of successful flights), buts the number of flights for these launch site significantly less for final conclusions.

Payload vs. Launch Site

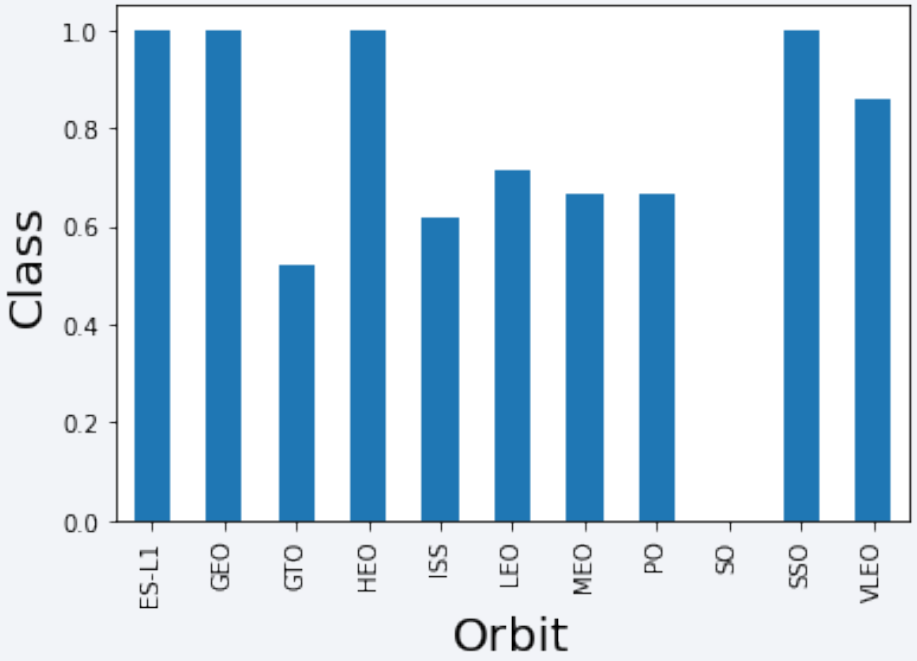


- For launch site CCAFS SLC 40, we have:
  - At payload mass < 8000kg, 25 (~45.45 %) successful flights are successful and 21(~38.18 %) failed flights
  - At payload mass > 8000 kg, 8 (~14.55 %) successful flights are successful and 1 (~1.82 %) failed flight
- For launch site VAFB SLC 4E, we have:
  - At payload mass < 8000kg, 4 (~30.77 %) successful flights are successful and 2 (~15.38 %) failed flights
  - At payload mass > 8000kg, 6 (~46.16 %) successful flights are successful and 1 (~7.69 %) failed flight
- For launch site KSL LC 39A, we have:
  - At payload mass < 8000kg, 11 (=50.00 %) successful flights are successful and 4 (~18.18 %) failed flights
  - At payload mass > 8000kg, 6 (~27.27 %) successful flights are successful and 1 (~4.55 %) failed flight

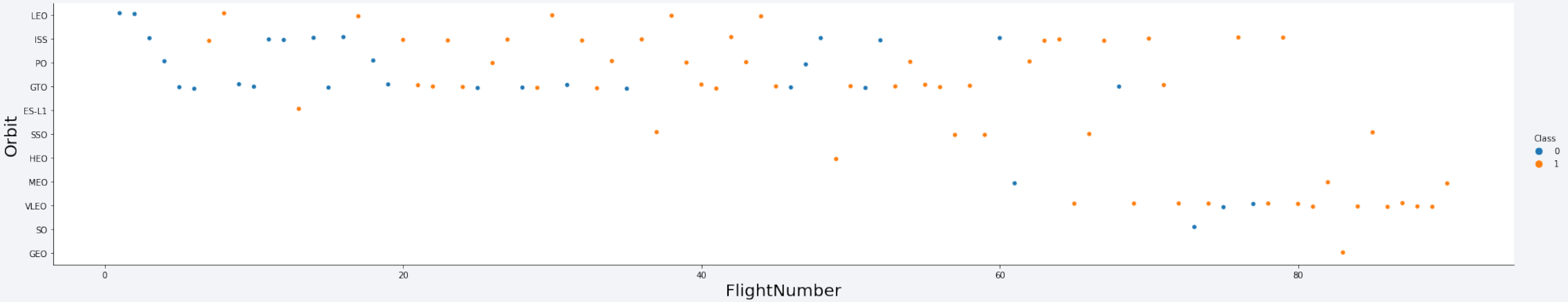
- Number of unsuccessful flights is higher at payload mass < 8000 kg. At payload mass > 8000 kg, we have 1 failed flight for all launch sites

Success Rate vs. Orbit Type

- The highest success rate of space flights occur on orbits:
  - ES-L1 (=100%)
  - GEO (=100%)
  - HEO (=100%)
  - SSO (=100%)
- And
  - VLEO (~85.71%)
  - LEO (~71.43%)

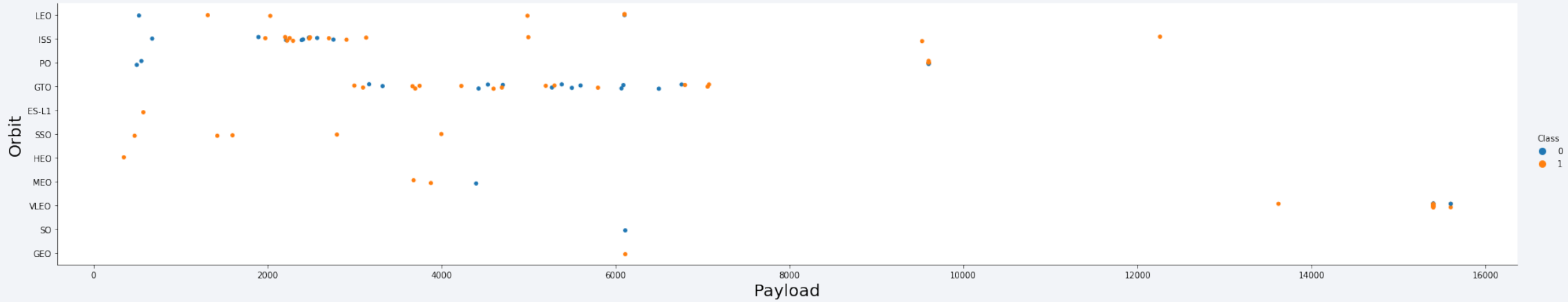


Flight Number vs. Orbit Type



- ES L1 & GEO & HEO 1 successful space flight (success rate = 100 %);
- GTO 14 successful and 13 failed space flights (success rate ~ 51.85 %);
- ISS 13 successful and 8 failed space flights (success rate ~ 61.9%);
- LEO 5 successful and 2 failed space flights (success rate ~ 71.43%);
- MEO 2 successful and 1 failed space flight (success rate ~ 66.67%);
- PO 6 successful and 3 failed space flights (success rate ~ 66.67%);
- SO 1 failed space flight (success rate = 0 %);
- SSO 5 successful space flights (success rate = 100 %); ←---------- The best case
- VLEO 12 successful and 2 failed space flights (success rate ~ 85.71%)

Payload vs. Orbit Type



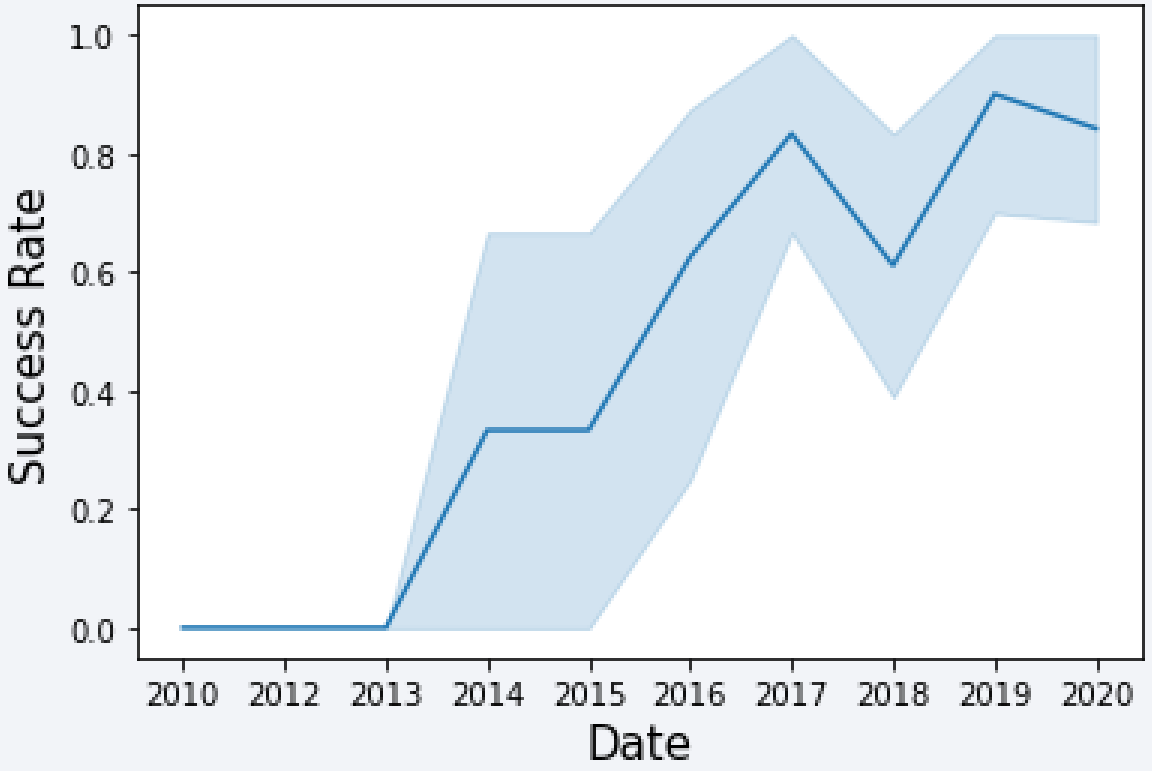
HEO, ES-L1& GEO :
•1 successful flight with payload 350 kg, 570 kg and 6,105 accordingly.
GTO :
•14 successful flights with payload in a range 3,000…7,076 kg;
•13 unsuccessful flights with payload in a range 3,170…6,761 kg.
ISS :
•13 successful flights with payload in a range 1,977…12,259 kg;
• 8 unsuccessful flights with payload in a range 677…2,760 kg.
LEO:
•5 successful flights with payload in a range 1,316…6,105 kg;
•2 unsuccessful flights with payload 525 and 6,105 kg.
MEO :
•2 successful flights with payload 3,681 kg and 3,880 kg accordingly;
•1 unsuccessful flight with payload 4,400 kg.
PO :
•6 successful flights with payload mass 9600 kg;
•3 unsuccessful flights with payload in a range 500…9,600 kg.
SO:
•1 unsuccessful flight with payload 6,105 kg.

SSO :
•5 successful space flights with payload in a range 475…1,425 kg;
VLEO:
•12 successful flights with payload between 13,620 and 15,400 kg;
•2 unsuccessful flights with payload between 15,400 and 15,600 kg.

- There is no relation between payload and success rate to orbit GTO;
- One success launch to the orbits HEO, ES-L1, GEO and SO;
- ISS orbit has the widest range of payload mass and not bad rate of success (**~62%**);
- SSO orbit has the most success rate (=**100 %**) but not wide range of payload mass;
- VLEO orbit has the biggest payload mass 15,400 kg and a good success rate (~ **85.71 %**)

Launch Success Yearly Trend

- During 2010-2013 no success flights;
- From 2013 we observe an increase in success rate.



All Launch Site Names

- Found four unique launch sites in space mission by applying SQL SELECT DISTINCT statement to return only distinct (different) “ launch_site ” values from dataset orders by first column

```
SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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`
- All 5 records show Cape Canaveral launches

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

- Total payload carried by boosters from NASA
- Result obtained with summing all payloads where customer contain “CRS”, which corresponds to NASA

```
SELECT SUM(PAYLOAD_MASS__KG) AS Total_Payload LAUNCH_SITE FROM SPACEXTBL
WHERE Customer LIKE ‘%CRS%’;
```

Total Payload (kg)
48213

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1
- The result obtained with AVG function applying to filtered data by the booster version in order to calculate the average payload mass

```
SELECT AVG(PAYLOAD_MASS__KG) AS AVG_Payload FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1';
```

Avg Payload (kg)
2928.4

First Successful Ground Landing Date

- First successful landing outcome on ground pad
- The result obtained with MIN function applying to filtered data by successful landing outcome on ground pad in order to identify the first occurrence

```
SELECT MIN(DATE) AS First_Success FROM SPACEXTBL
WHERE Landing_Outcome= 'Success (ground pad)';
```

First success date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- There are five results according to the filters above

```
SELECT DISTINCT BOOSTER_VESRION FROM SPACEXTBL
WHERE Landing_Outcome= 'Success (drone ship)' AND PAYLOADF_MASS__KG BETWEEN 4000 AND 6000;
```

Booster Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1038.1
F9 FT B1038.2

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes
- Result obtained by grouping mission outcomes and counting records for each group

```
SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL
GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

Mission Outcome	Quantity
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass
- Ordered booster versions obtained from SPACEXTBL with applying MAX function to find maximum payload mass and ORDER BY keywords to sort the results

```
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_ )
FROM SPACEXTBL)
ORDER BY BOOSTER_VERSION;
```

Booster Version
F9 B4 B1041.2
F9 B4 B1041.1
F9 B5 B1049.2
F9 B5 B1048.1
F9 FT B1036.2
F9 FT B1029.1
F9 FT B1036.1

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Result obtained from SPACEXTBL with applying date function to filtering landing outcomes for in year 2015 and LIKE operator in a WHERE clause to find for a “Fail” keywords in a landing outcome column.

```
SELECT date,booster_version,launch_site,landing_outcome FROM SPACEXTBL
WHERE strftime("%Y",Date)="2015” AND Landing_Outcome LIKE "%Fail%"
```

Date	Booster Version	Launch Site	Landing Outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

- 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
- Result obtained from SPACEXTBL with applying date function to filtering landing outcomes between the date 2010-06-04 and 2017-03-20 and GROUP BY statement in a WHERE clause to group landing outcome rows that have the same values into summary rows and ORDER BY keyword to sort to sort the quantity set in descending order.

Landing Outcome	Quantity
No attempt	9
Failure (drone ship)	5
Success (drone ship)	4
Controlled (ocean)	3
Failure (parachute)	2
Success (ground pad)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

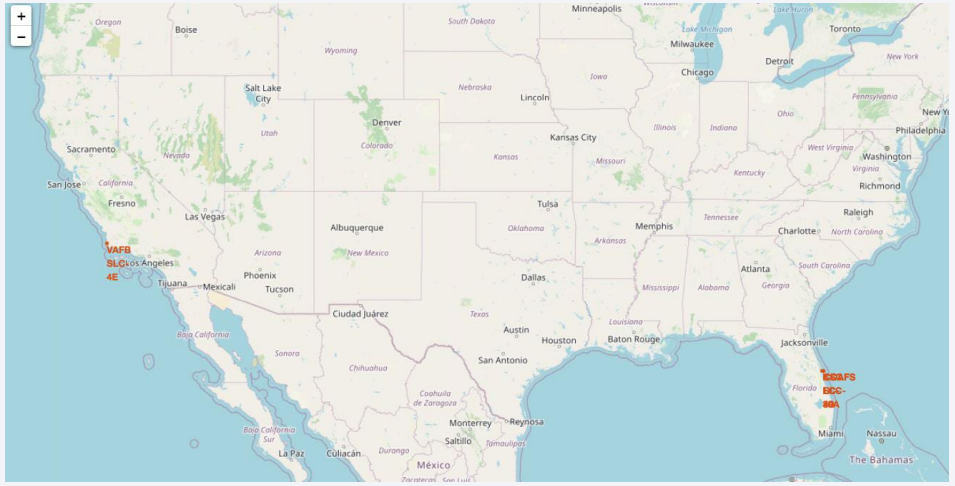
```
SELECT Landing_Outcome,COUNT(*)AS Quantity FROM SPACEXTBL
WHERE strftime("%Y%m%d",date) BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER by Quantity DESC
```

Section 3

Launch Sites Proximities Analysis

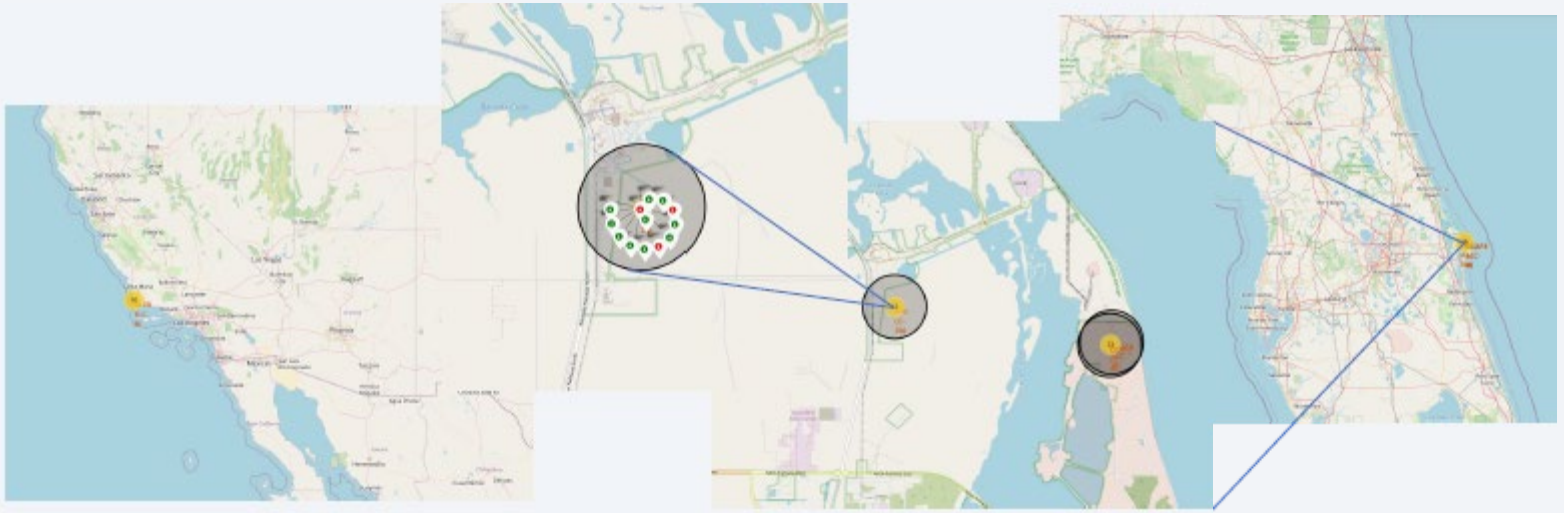
All Launch Sites

- All launch sites in proximity to the Equator line and to the coast. Location near equator gives an additional natural boost due to the rational speed of Earth that helps save on fuel and boosters. Reason of location near coast is that if something goes wrong during the ascent, the debris will fall into an ocean instead of a densely populated area.



Launch outcomes for each site

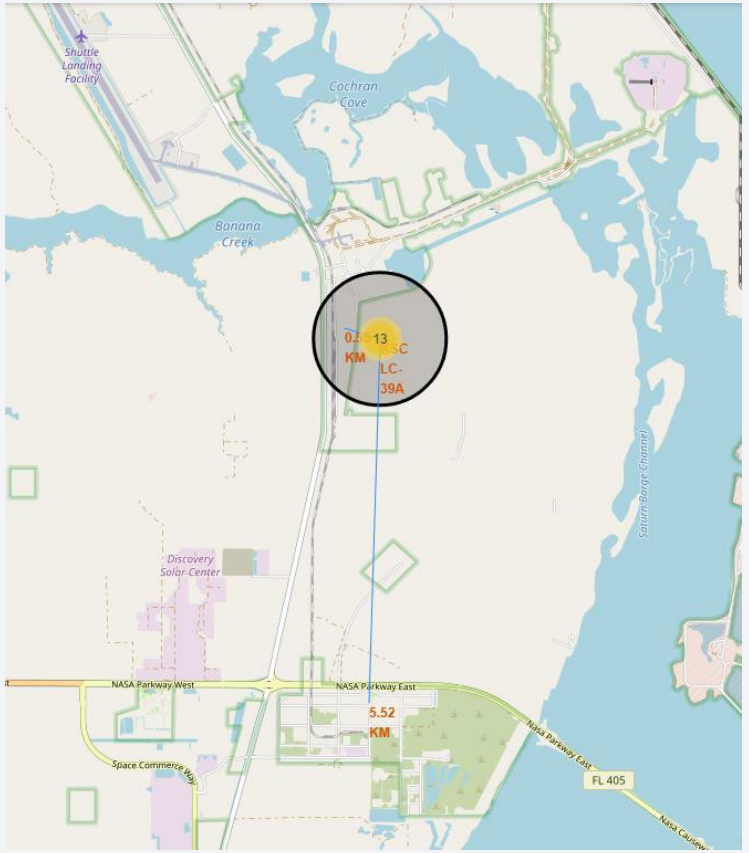
- This is an example of launch outcomes for KSC LC-39A launch site. Green markers indicate successful launch outcomes and red ones indicate failure outcomes on the map.

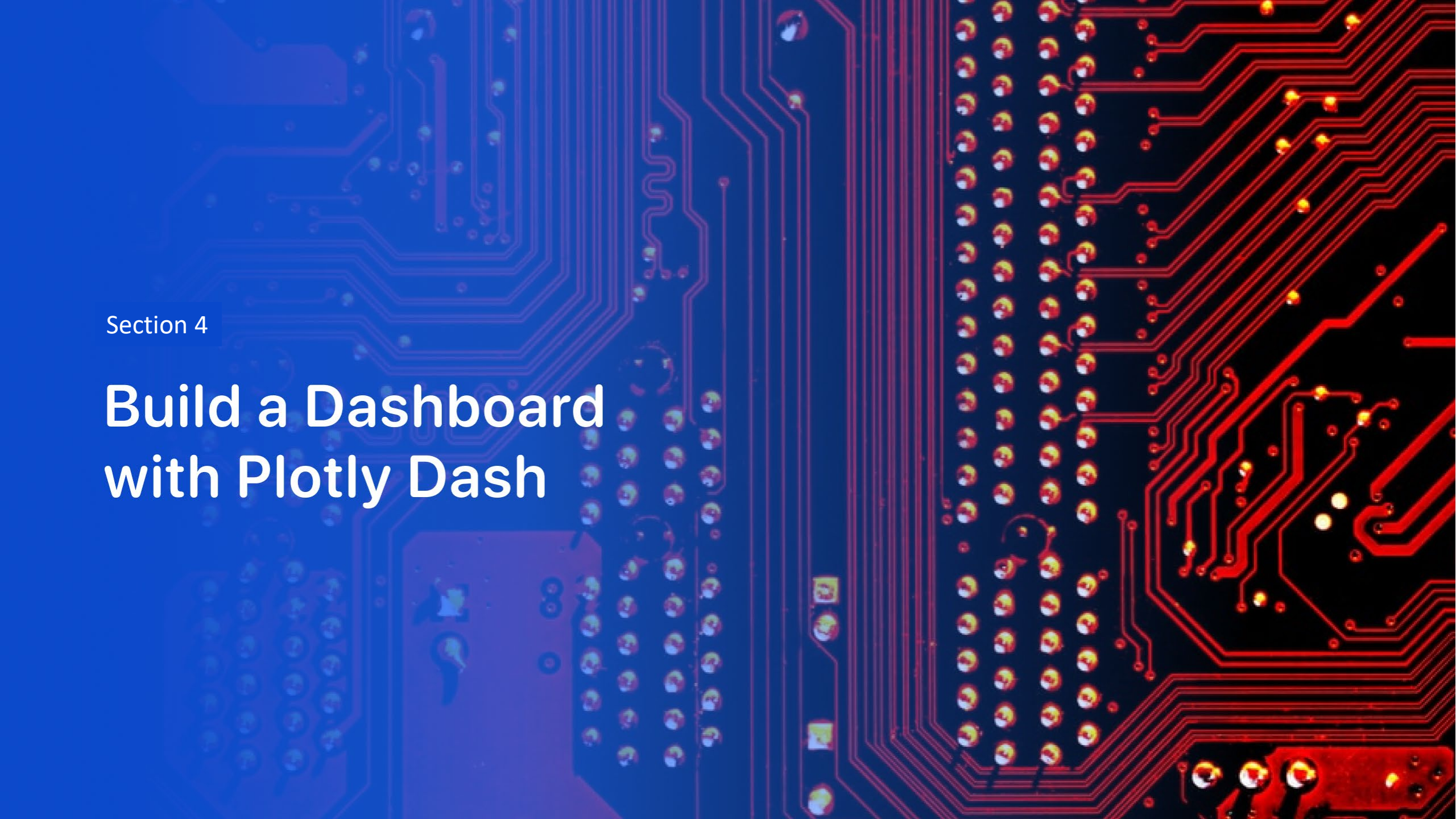


Logsitics

- Example of location for KSC LC-39A to find that
  - Launch site is near to railway
  - Launch site is near to highways
  - Launch site is near to coastline
  - Launch site is relatively far from cities

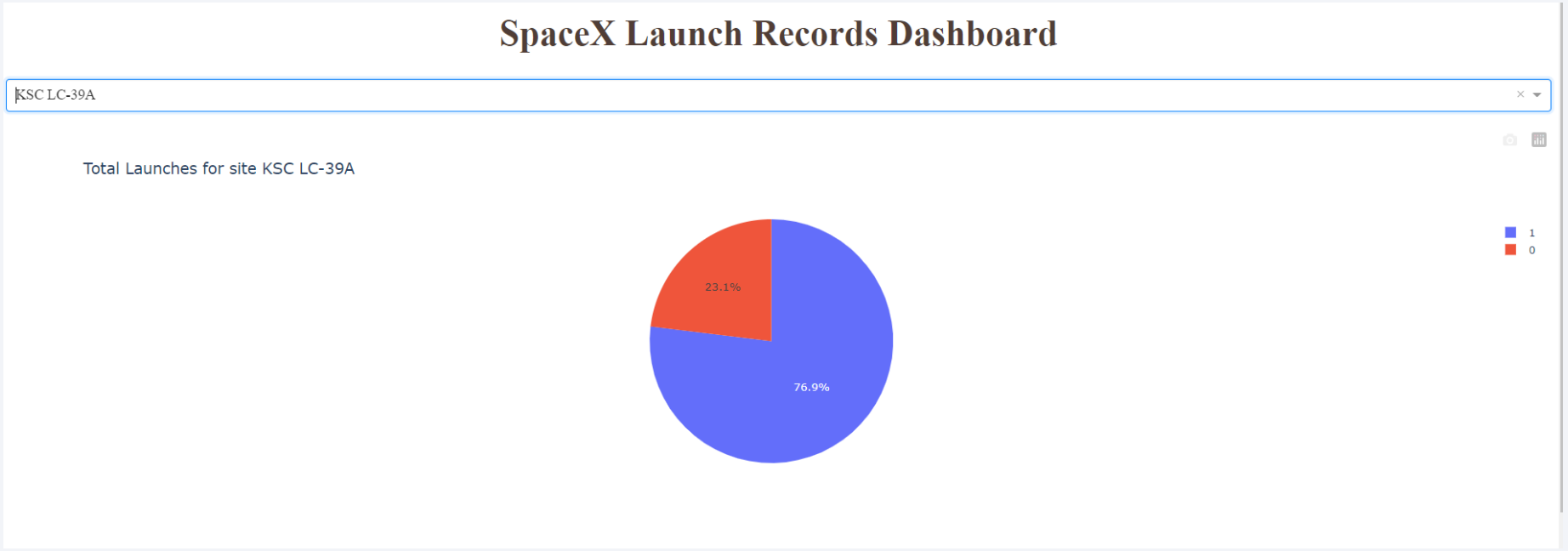
- Thus KSC LC-39A launch site has good logistics aspects.




Section 4
Build a Dashboard with Plotly Dash

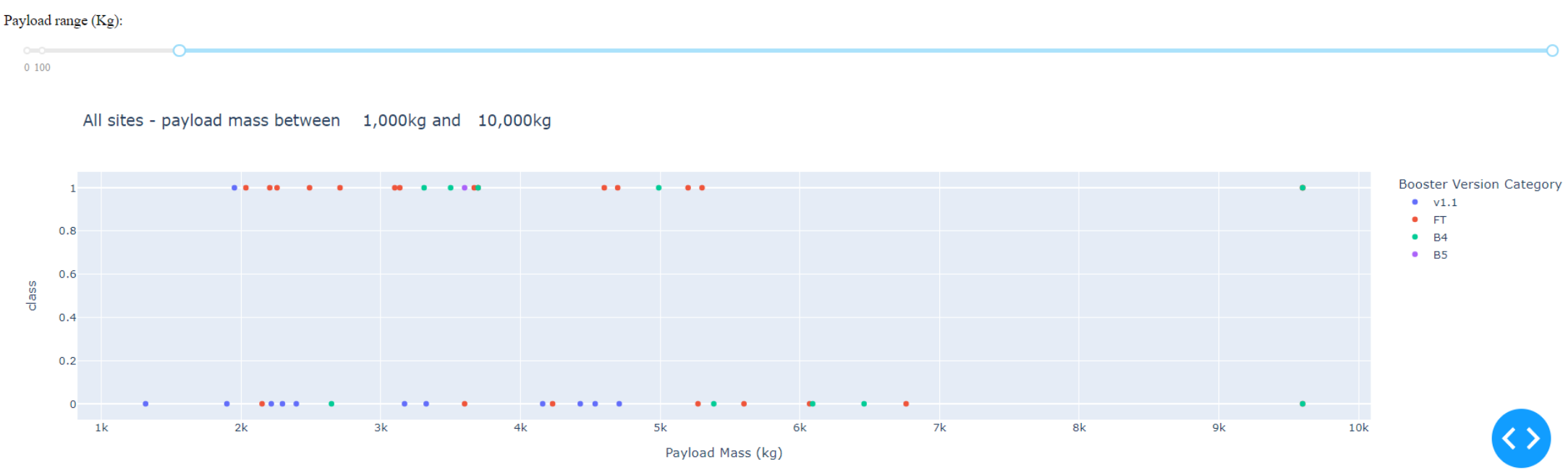
Total launches for site KSL LC-39A

- 76.9 % launches are successful.



Payload vs Launch Outcome for all sites

- Payload range 2000 – 6000 kg and FT booster have the largest success rate.

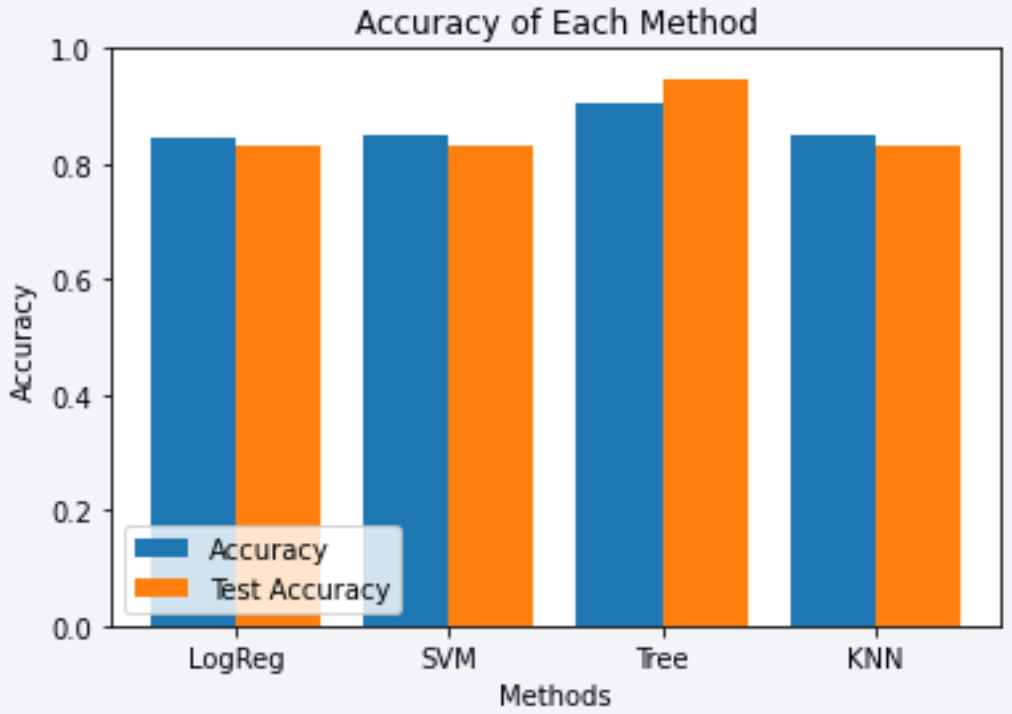


Section 5

Predictive Analysis (Classification)

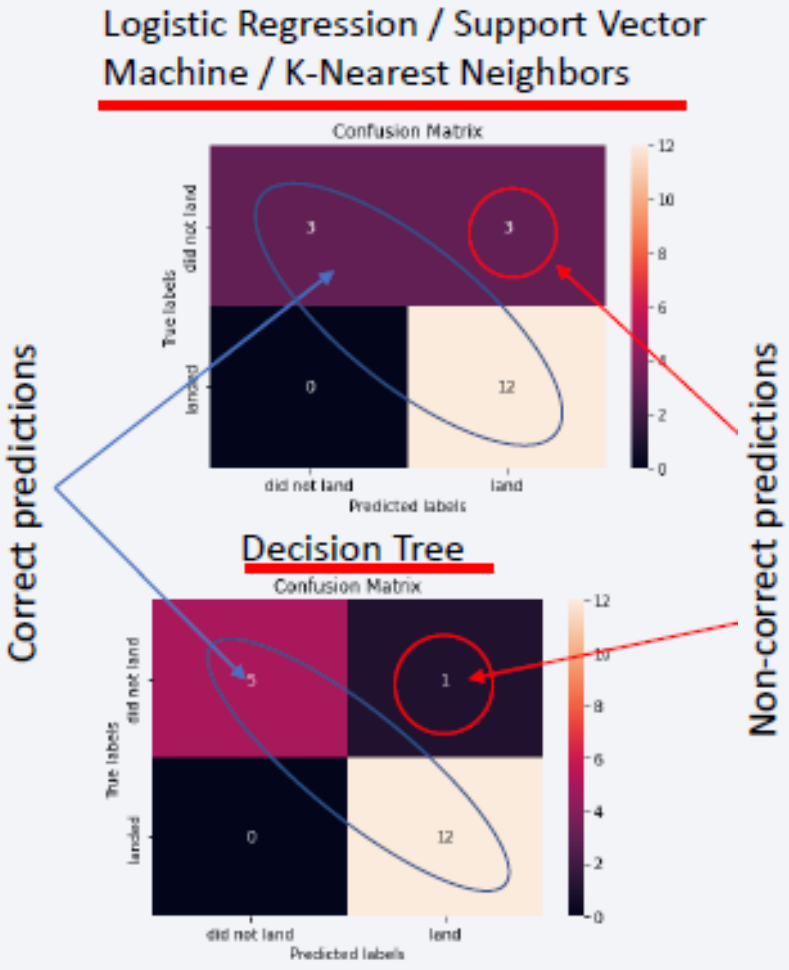
Classification Accuracy

- Four classification models were tested
- Logistic Regression, Support Vector, Machine, K Nearest Neighbors models show the same accuracy ~ 85.0%
- The Decision Tree Classifier show the highest accuracy over 90.0%



Confusion Matrix

- In confusion matrix, the Log Regression / Support Vector Machine / K Nearest Neighbors models judged that from 6 unsuccessful launches (did not land) 3 were successful and instead of 12 successful launches (land), it predicted that 15 have landings (12 correct predicted + 3 non correct predicted).
- Decision Tree (DT) model judged 1 wrong prediction. That is from 6 unsuccessful landings DT model predict one as successful and instead of 12 landings, it predicted that 13 have landings (1 was wrong prediction). All correct predictions are in the diagonal of the table, so it is easy to visually inspect the table for prediction errors, as values outside the diagonal will represent them . Thus, Decision Tree model proves its accuracy by showing the big numbers of true positive values (12) and true negative (5) compared to the false ones.



Conclusions

- Different data sources were analyzed, refining conclusions along the process;
- The KSC LC-39A launch site has good logistics aspects and more then 76% of successful flights;
- At payload mass > 8000 kg, we have higher successful flights for all launch sites;
- VLEO orbit has the biggest payload mass 15400 kg and a good succuess rate (more than 85%);
- Starting from 2013, most of mission outcomes are successful due to the improvements of rockets;
- Decision Tree Classifier has shown better accuracy compared to Logistic Regress, Support Vector Machine, K-Nearest Neighbors models to predict successful landings.

Appendix

- https://github.com/wttz1212/Applied-Data-Science-Capstone

Thank you!