IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Isabella Guerra

5th of May 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results/Insights
- Conclusion
- Further Work Recommendations
- Appendix

Executive Summary

Predicting if Space X Falcon 9 first stage will land successfully or not

- Methodologies

Data Collection: API and web scraping

Data Wrangling

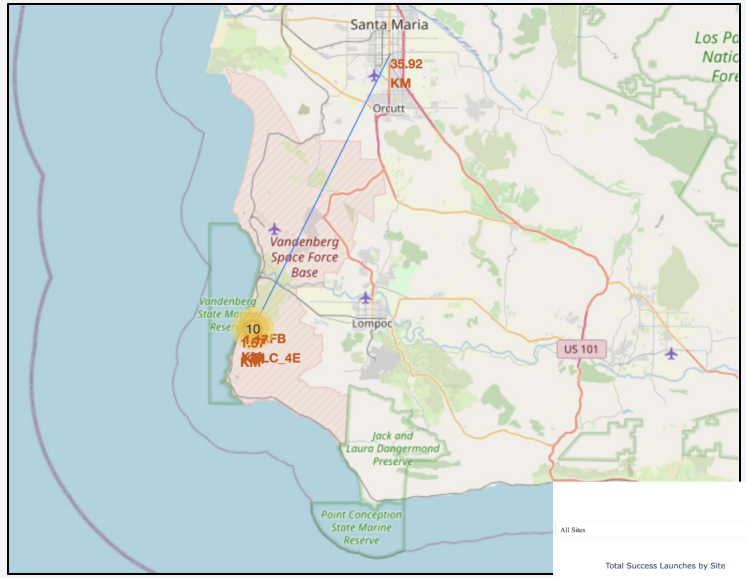
Exploratory Data Analysis with visualization and SQL

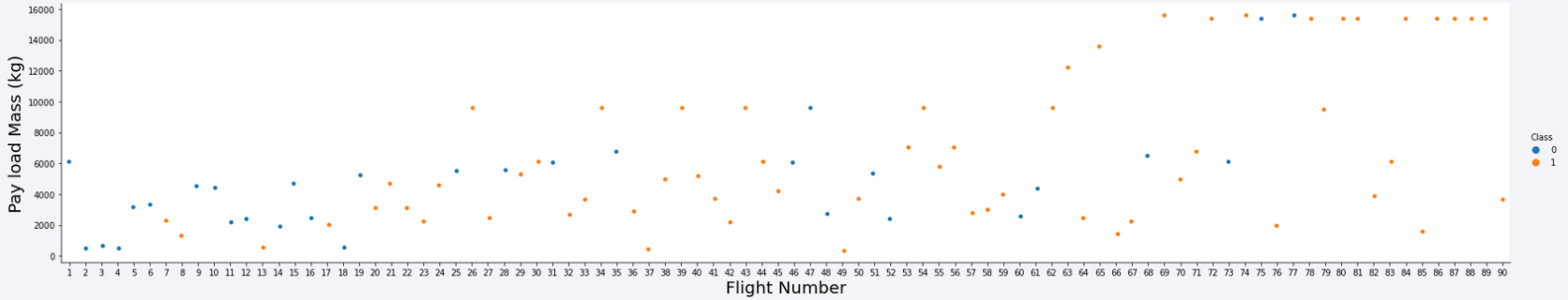
Interactive Visual Analytics with Folium and Plotly Dash

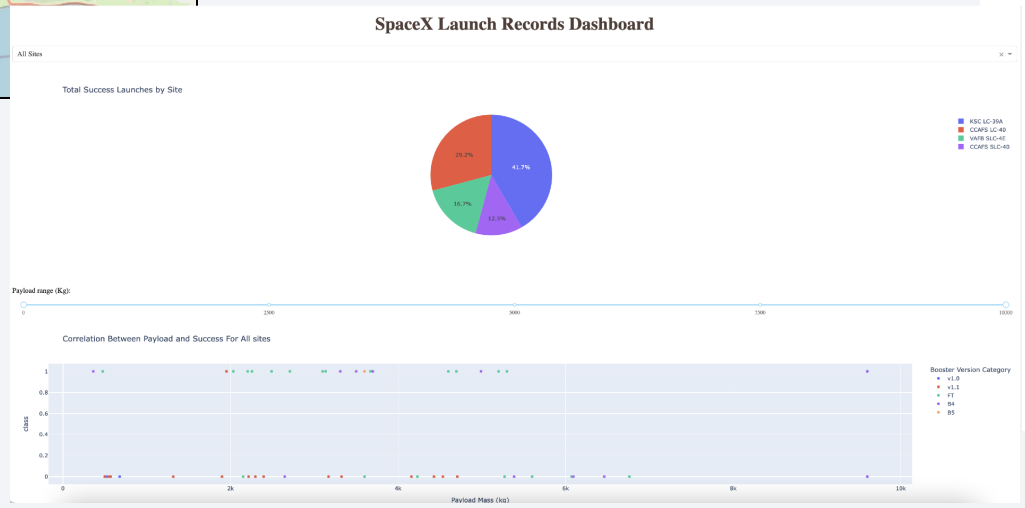
Predictive Analysis using classification models: LR, SVM, Decision Tree, KNN

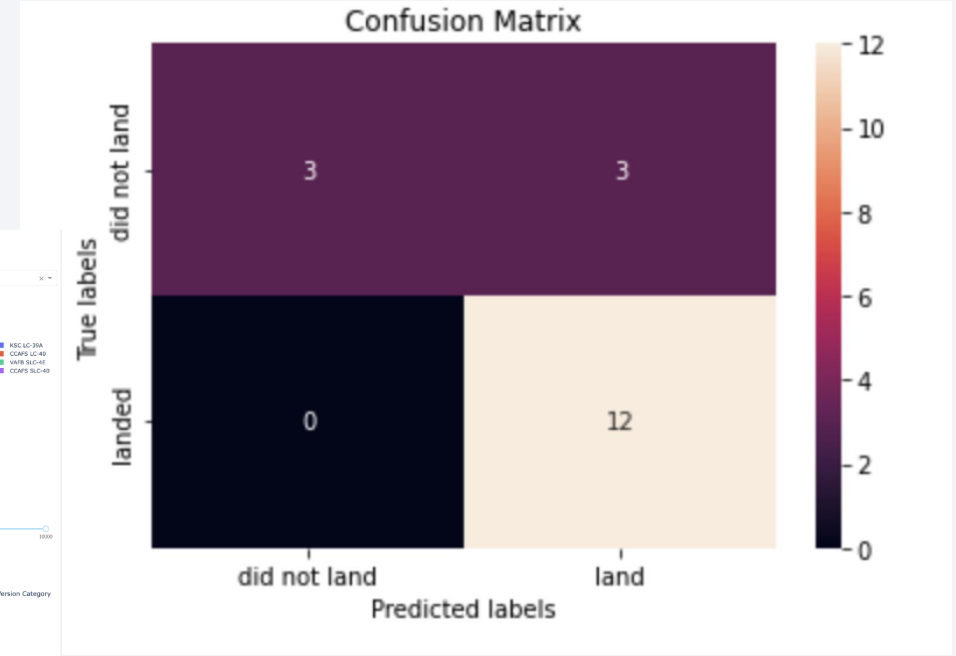
Executive Summary

- Results









Introduction – Project Background

- The cost of launching a Falcon 9 rocket is reported by aerospace manufacturers as more than 165million $
- SpaceX advertises Falcon 9 rocket launches for a total cost of 62million $, a much lower rate compared to competitors
- The saving is related to the fact the SpaceX is counting on reusing the first stage

If we can determine if the first stage will land, we can determine the total cost of a Falcon 9 launch

Introduction – The problem

Will the First Stage of Falcon 9 land successfully?

What are the factors implicated in a successfully landing?

- Earth Space Location?
- Payload Mass?
- Launch Locations?
- Booster Version?
- Time of launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Space X API & Wikipedia Web Scraping
- Perform data wrangling:
  - Data was cleaned and adapted for the analysis
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models:
  - Logistic Regression, Decision Tree, Support Vector Machine, K Nearest Neighbour were tested

Methodology - Data Collection

Data Sources:

1) Space X API

- Historical launches data for Falcon 1 and Falcon 9 was requested from SpaceX API
- Only data for Falcon 9 was used
- The data was cleaned in order to perform the analysis

2) Wikipedia Web Scraping

- Falcon 9 launch records were extracted from an HTML Wikipedia table
- The data was cleaned in order to perform the analysis

Data Collection – SpaceX API

Workflow:

1) Requesting the data

Now let's start requesting rocket launch data from SpaceX API with the following URL:

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

2) Response decoded as a Json and turned into Pandas dataframe

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
# Use json_normalize meethod to convert the json result into a dataframe
response=response.json()
data=pd.json_normalize(response)
```

3) The API was used to get information about the launches, the data was then stored in lists and a new panda dataframe was created

Then, we need to create a Pandas data frame from the dictionary launch_dict.

```
# Create a data from launch_dict
launch_dict_df=pd.DataFrame(launch_dict)
```

Data Collection – SpaceX API

4) Final result after some data manipulations and having excluded Falcon 1 launches

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitu
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.5773
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.5773
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.5773
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.6108
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.5773

GitHub Notebook URL:

https://github.com/Isabella8833/bug-free-journey/blob/master/Capstone%20Project%20-%20Data%20Collection%20API.ipyn

Data Collection – Web Scraping

Workflow:

1) HTTP GET method was performed to request the Falcon 9 launch HTML page

2) Beautiful Soup object from the HTML response was created

3) All HTML tables in the Wikipedia page were found, the one containing the relevant information about Falcon 9 was extracted

Data Collection – Web Scraping

4) A data frame was created after parsing the HTML records into a dictionary

After you have fill in the parsed launch record values into `launch_dict`, you can create a dataframe from it.

```
#df=pd.DataFrame(launch_dict)
df= pd.DataFrame({ key:pd.Series(value) for key, value in launch_dict.items() })
```

5) The dataframe was then exported to a CSV file

Following labs will be using a provided dataset to make each lab independent.

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

GitHub Notebook URL:

https://github.com/Isabella8833/bug-free-journey/blob/master/Capstone%20Project%20-%20Data%20Collection%20Web%20Scraping.ipynb

Methodology - Data Wrangling

```
df.dtypes
```

```
FlightNumber          int64
Date                 object
BoosterVersion       object
PayloadMass         float64
Orbit                object
LaunchSite           object
Outcome              object
Flights               int64
GridFins               bool
Reused                 bool
Legs                   bool
LandingPad           object
Block               float64
ReusedCount           int64
Serial               object
Longitude           float64
Latitude            float64
```

Missing values were identified

Identify and calculate the percentage of the missing values in each attribute

```
df.isnull().sum()/df.count()*100
```

```
FlightNumber      0.000
Date              0.000
BoosterVersion    0.000
PayloadMass       0.000
Orbit             0.000
LaunchSite        0.000
Outcome           0.000
Flights           0.000
GridFins          0.000
Reused            0.000
Legs              0.000
LandingPad       40.625
Block             0.000
ReusedCount       0.000
Serial            0.000
Longitude         0.000
Latitude          0.000
```

Columns data types were double-checked

Number of occurrences was calculated for different parameters (Orbit, Launch Site..)

```
# Apply value_counts() on column LaunchSite
df["LaunchSite"].value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
```

A binary landing outcome was created for each launch entry (0=failed landing, 1=successful landing) and saved in a new column called "Class"

```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise

landing_class=[]

for i in df["Outcome"]:
    if i in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)

landing_class
```

GitHub Notebook URL:

https://github.com/Isabella8833/bug-free-journey/blob/master/Capstone%20Project%20-%20Data%20Wrangling.ipynb

Methodology - EDA with Data Visualization

The relationships between the parameters were investigated to understand if they influence the first stage landing outcome

Plots and charts were built

```
sns.catplot(y="PayloadMass", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Pay load Mass (kg)",fontsize=20)
plt.show()
```

Dummy variables were created for categorical columns to be included in the subsequent analysis

```
features_one_hot = pd.get_dummies(features, columns=['Orbit','LaunchSite','LandingPad','Serial'])
features_one_hot.head()
```

GitHub Notebook URL:

https://github.com/Isabella8833/bug-free-journey/blob/master/Capstone%20Project%20-%20EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

Different SQL queries were performed in order to analyse the data:

- Analysis of first stage landing outcomes and missions outcomes
- Analysis of booster versions and payload masses
- Analysis of records in specific time periods

```
%%sql

Select BOOSTER_VERSION from SPACEXDATASET where "Landing _Outcome" like '%Success (drone ship)%' and PAYLOAD_MASS__KG_
between 4000 and 6000;
```

```
 * ibm_db_sa://stf29843:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

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

```
%%sql

Select "Landing _Outcome", BOOSTER_VERSION, LAUNCH_SITE, DATE from SPACEXDATASET
 where "Landing _Outcome" like '%Failure (drone ship)%'
    and Year(Date)=2015;
```

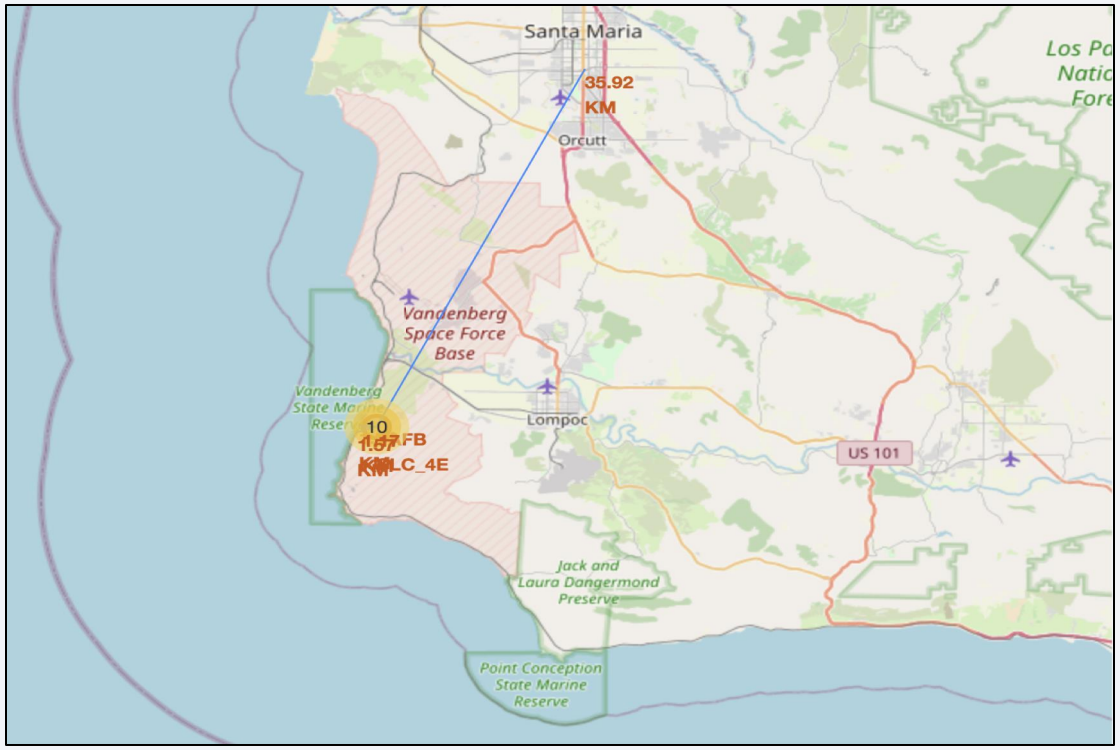
```
 * ibm_db_sa://stf29843:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8l
Done.
```

Landing _Outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

GitHub Notebook URL:

https://github.com/Isabella8833/bug-free-journey/blob/master/Capstone%20Project%20-%20EDA%20with%20SQL_GitHub.ipynb

Methodology - Build an Interactive Map with Folium

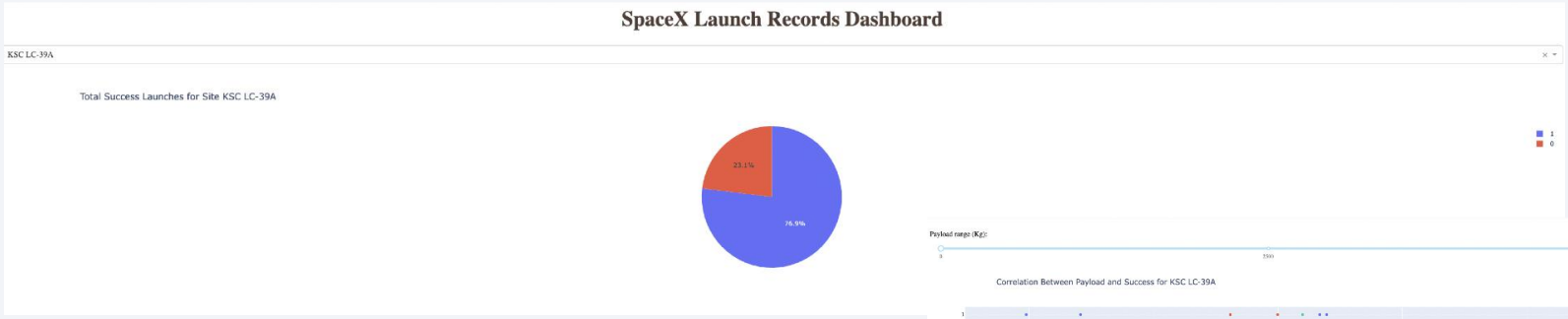


- Markers, circles and lines were created and added to a Folium map
- Those objects were used in to analyze the launch sites locations
- The goal of the analysis was understanding if launch site location has an effect on the success or failure of a launch

GitHub Notebook URL:

https://github.com/Isabella8833/bug-free-journey/blob/master/Capstone%20Project%20-%20Interactive%20Visual%20Anal.%20Folium.ipynb

Build a Dashboard with Plotly Dash

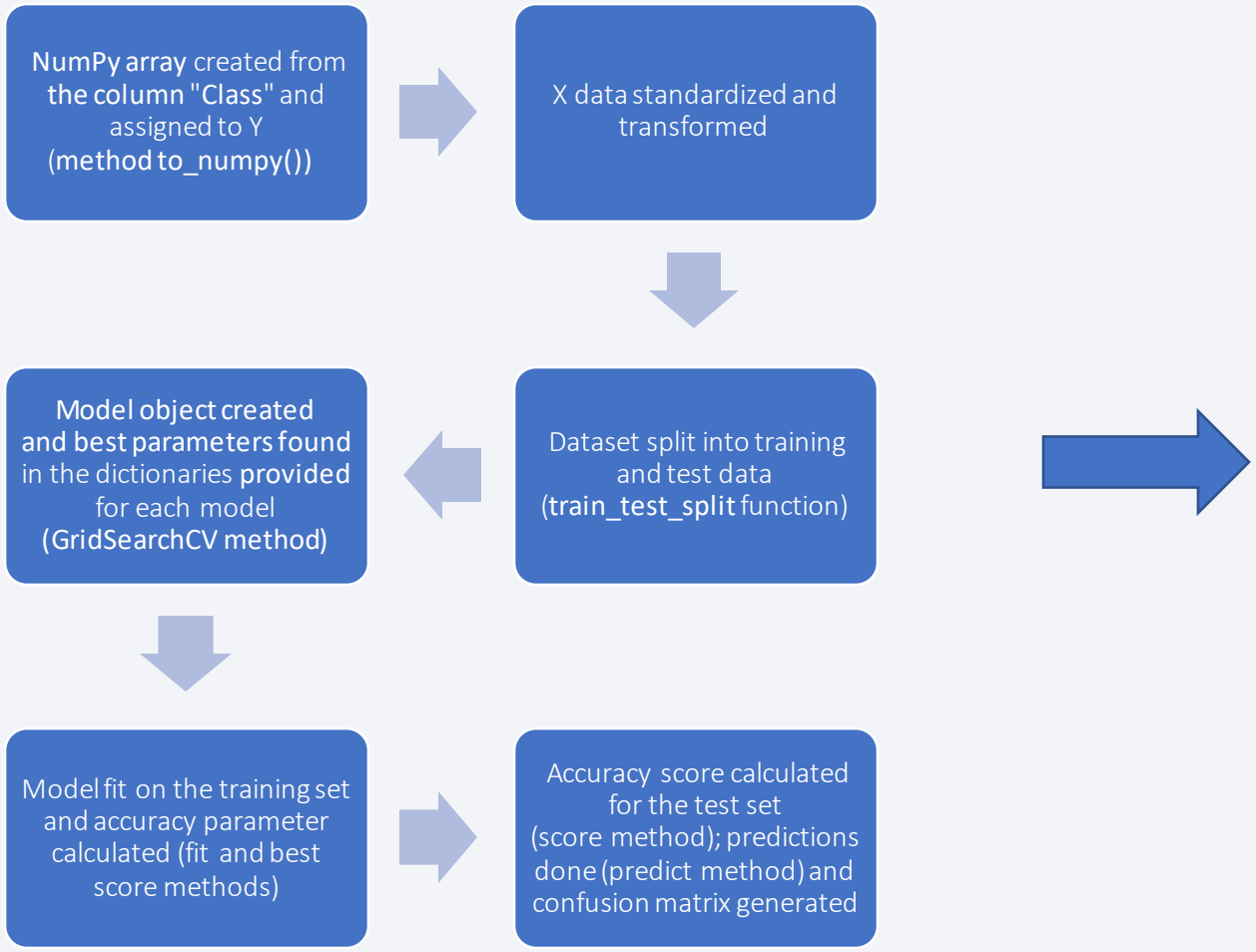


- Pie charts and scatter plots were built in an interactive dashboard using Plotly Dash
- Pie charts were used to investigate the rates of launch success and failure for all launch sites
- Scatter plots were used to understand the correlation between the payload mass and booster version with the success or failure of a launch

GitHub Notebook URL:

https://github.com/Isabella8833/bug-free-journey/blob/master/Capstone%20Project%20-%20Interactive%20Dashboard%20Ploty%20Dash.py

Methodology - Predictive Analysis (Classification)



Classification models tested:

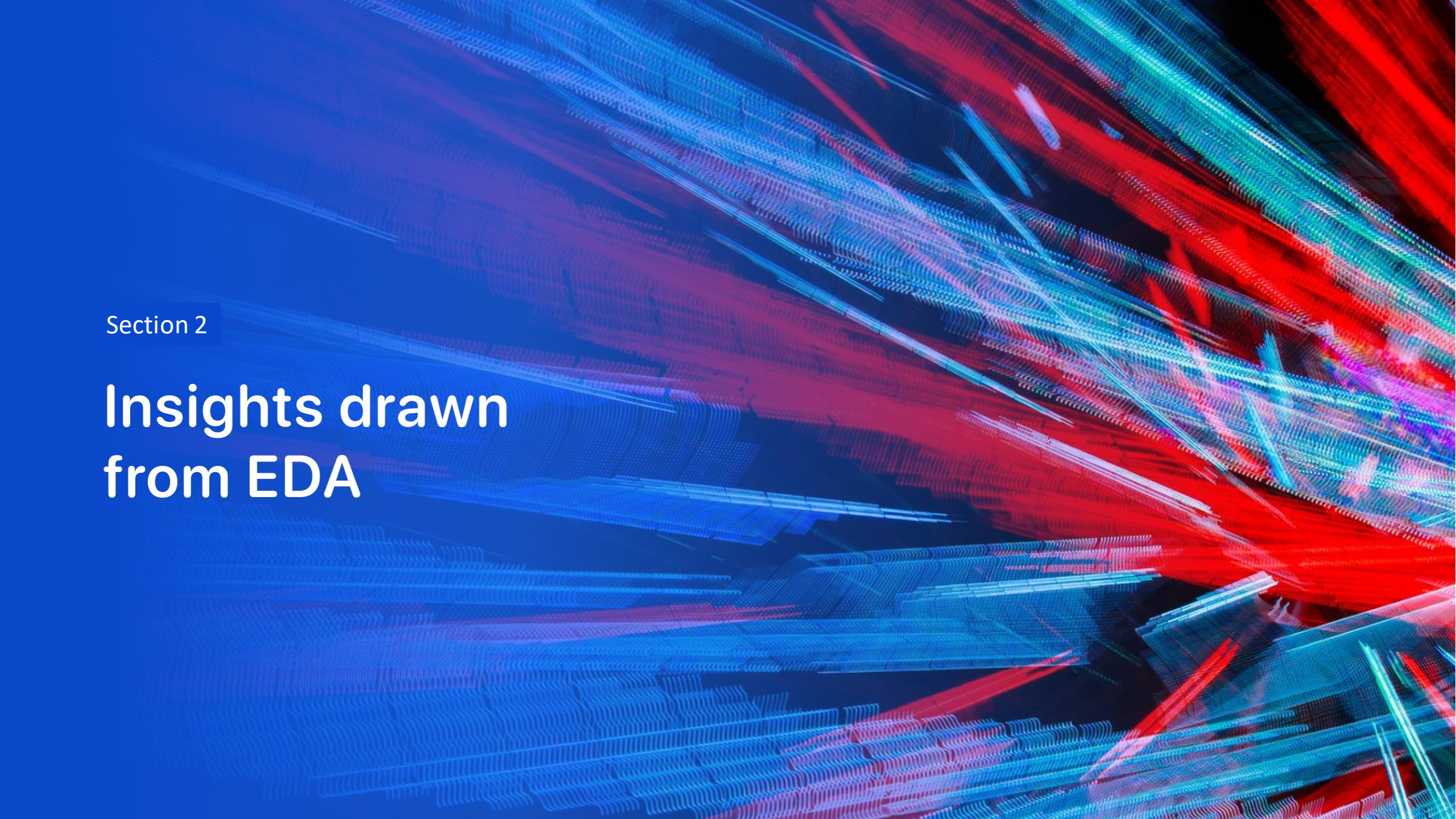
- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- K Nearest Neighbour (KNN)

GitHub Notebook URL:

https://github.com/Isabella8833/bug-free-journey/blob/master/Capstone%20Project%20-%20Machine%20Learning%20Prediction.ipynb

Section 2

Insights drawn from EDA

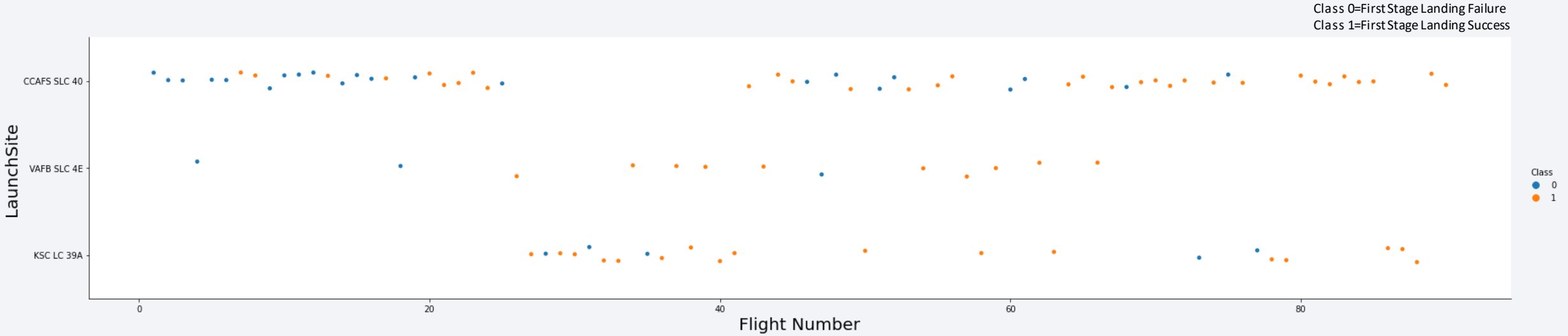


Section 2a

Insights drawn from EDA

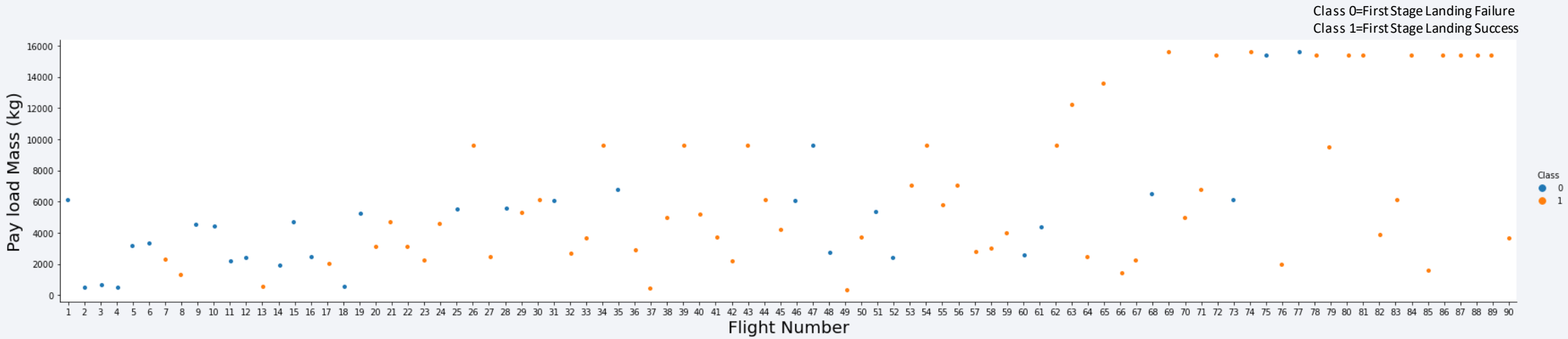
Results – Data Visualization

Flight Number vs. Launch Site



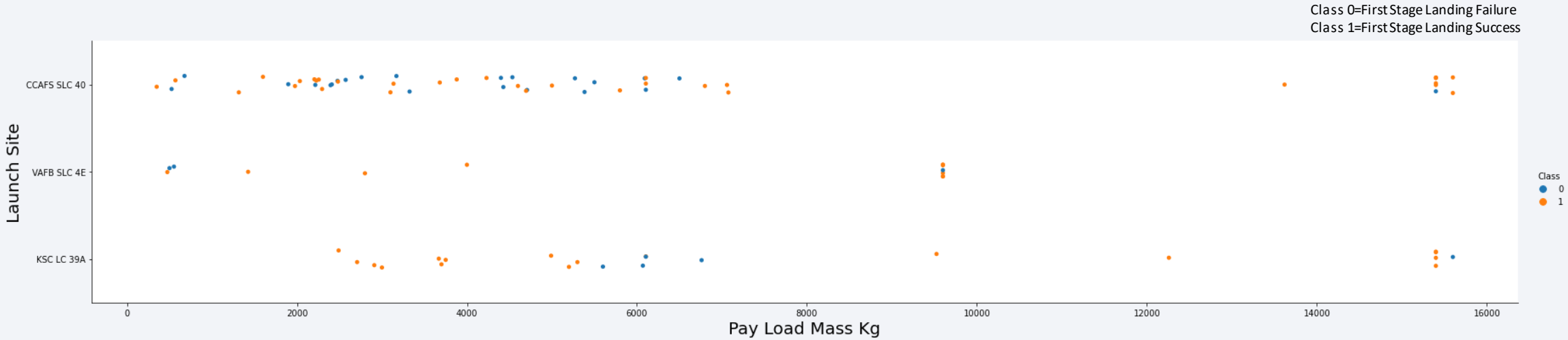
- The launch sites have different success rate
- As the flight number increases, the first stage is more likely to land successfully in all three sites
- Less continuous lunch attempts are performed in VAFB SLC 4E

Flight Number vs. Payload Mass



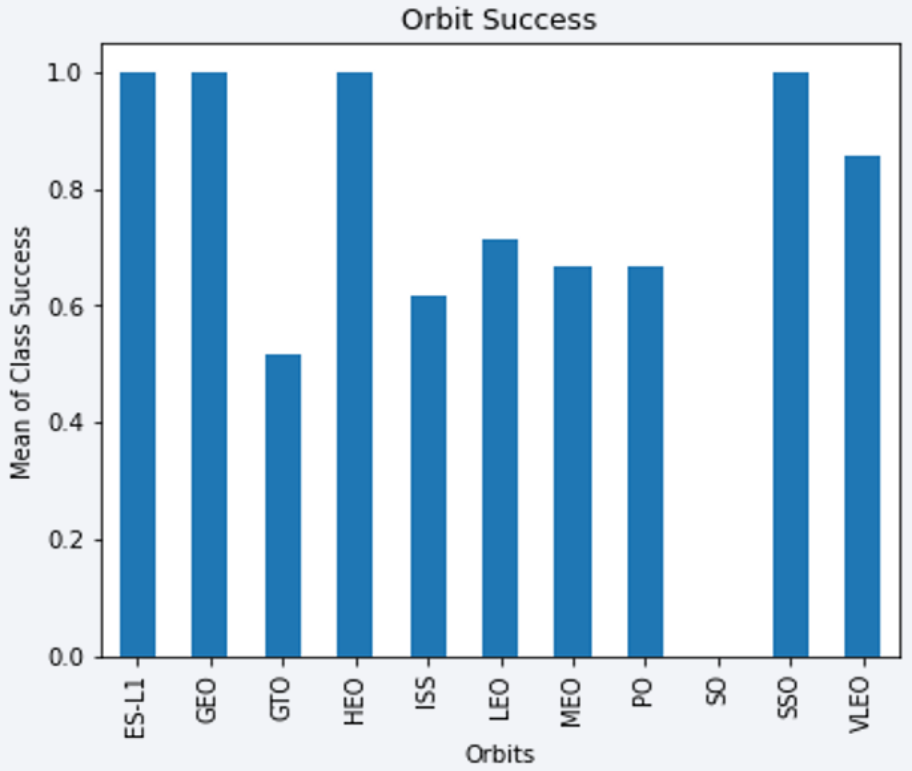
- The greater the weight of the payload and the less likely the first stage will return
- Massive payloads are more likely to return after many continuous launch attempts
- As the flight number increases, the first stage is more likely to land successfully

Payload Mass vs. Launch Site



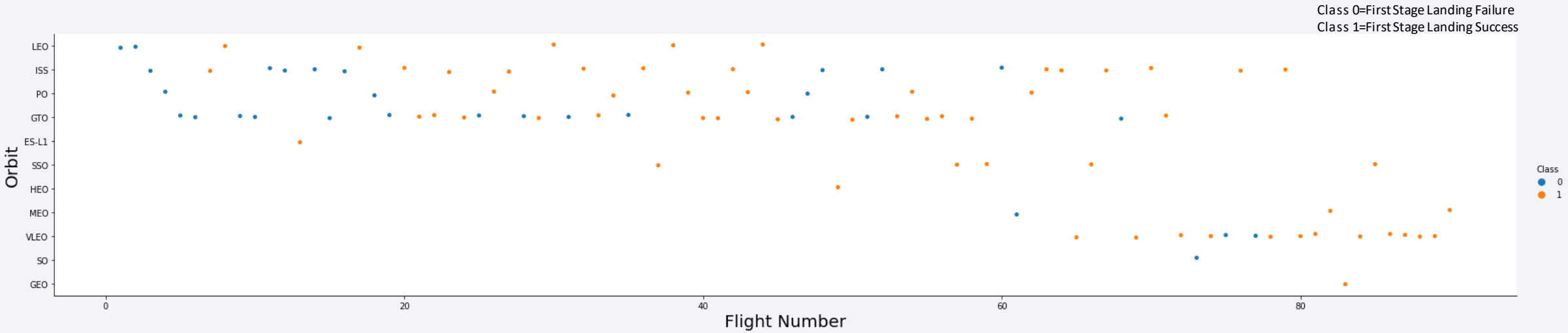
- It is a mixture of successes and failures launches with different payload masses for all three sites
- For VAFB-SLC launch site there are no rockets launched with high payload mass (more than 10000)

Success Rate vs. Orbit Type



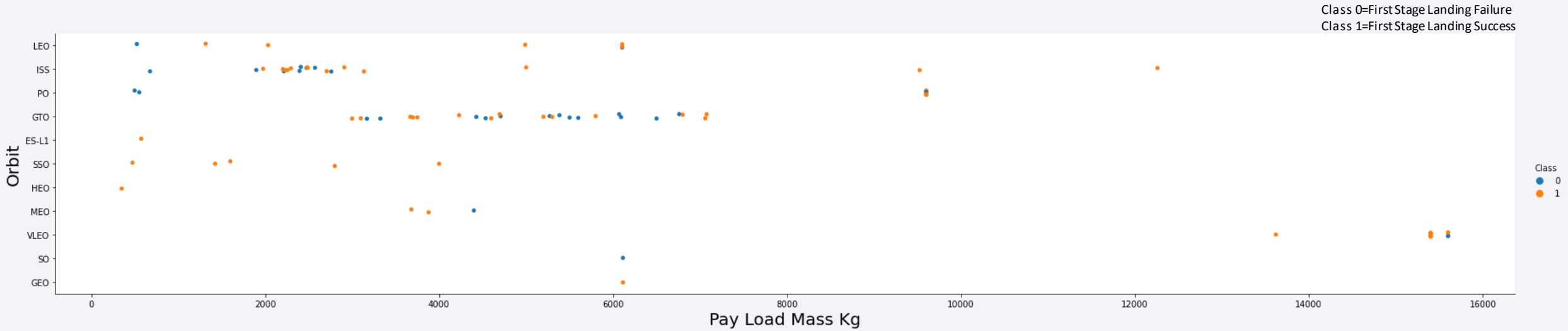
- The different orbits have different success rates of first stage landing
- ES-L1, GEO, HEO and SSO are the orbits with the highest success rate

Flight Number vs. Orbit Type



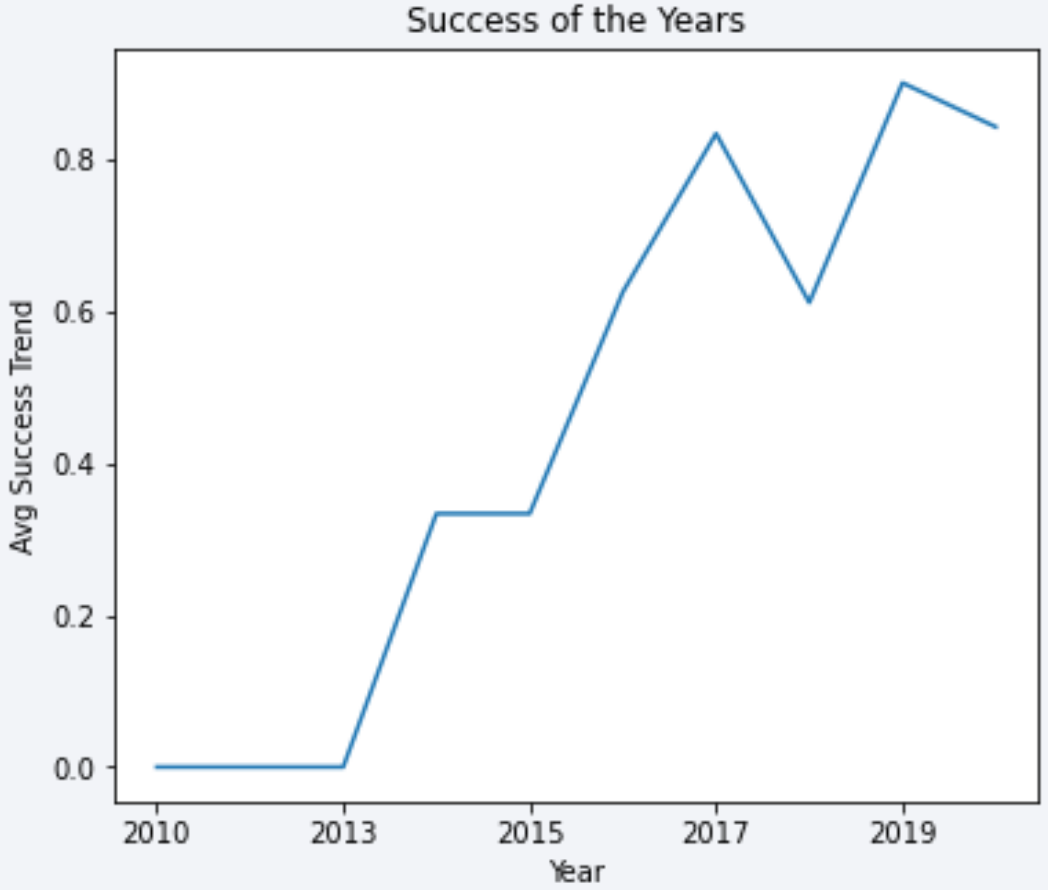
- For some orbits (ex. LEO, ISS), the higher the flight number and the more likely is the first stage to land successfully
- For other orbits (ex. GTO) the number of flight doesn't seem to affect the landing outcome
- For some other orbits there isn’t enough data to inform an opinion

Payload vs. Orbit Type



- For some orbits (ex. LEO, ISS), the higher the payload mass and the more likely is the first stage to land successfully
- For other orbits (ex. GTO), the payload mass doesn't see to affect the landing outcome (mix of successes and failures)
- For some other there isn’t enough data to inform an opinion

Launch Success Yearly Trend



- The success rate kept increasing from 2013
- The first stage is more likely to land successfully and then been reused nowadays compared 10 years ago

Section 2b

Insights drawn from EDA

Results - SQL

All Launch Site Names

```
%%sql

Select Distinct LAUNCH_SITE from SPACEXDATASET;
```

```
 * ibm_db_sa://stf29843:***@824dfd4d-99de-440d-99
Done.
```

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

The dataset had records for 4 different launch sites:

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

Launch Site Names Begin with 'CCA'

```
%%sql

Select * from SPACEXDATASET where LAUNCH_SITE like 'CCA%' limit 5;
```

```
 * ibm_db_sa://stf29843:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90108kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

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

5 records with launch site starting with 'CAA'

Total Payload Mass

```
%%sql
Select SUM(PAYLOAD_MASS__KG_) as Total_Payload_Mass_Kg from SPACEXDATASET where Customer='NASA (CRS)';
```

```
 * ibm_db_sa://stf29843:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomai
Done.
```

total_payload_mass_kg
45596

The total payload mass carried by boosters launched by NASA (CRS):

45596 kg

Average Payload Mass by F9 v1.1

```
%%sql

Select AVG(PAYLOAD_MASS__KG_) as Avg_Payload_Mass_Kg from SPACEXDATASET where Booster_Version='F9 v1.1';
```

```
 * ibm_db_sa://stf29843:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.
Done.
```

avg_payload_mass_kg
2928

The average payload mass carried by booster version F9 v1.1:

2928 kg

First Successful Ground Landing Date

```
%%sql

Select Min(Date) from SPACEXDATASET where "Landing _Outcome" like 'Success (ground pad)%';
```

```
 * ibm_db_sa://stf29843:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databa
Done.
        1
2015-12-22
```

The date when the first successful landing outcome in ground pad was achieved:

22nd of December 2015

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql

Select BOOSTER_VERSION from SPACEXDATASET where "Landing _Outcome" like '%Success (drone ship)%' and PAYLOAD_MASS__KG_
between 4000 and 6000;
```

```
 * ibm_db_sa://stf29843:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

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

The boosters which have success in drone ship and have payload mass between 4000 and 6000 are:

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

Total Number of Successful and Failure Mission Outcomes

```
%%sql

Select Distinct Mission_Outcome, Count(*) as Total_Num_FS from SPACEXDATASET
group by Mission_Outcome;
```

```
 * ibm_db_sa://stf29843:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb
Done.
```

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

Total number of successful missions: 99

Total number of failed missions: 2

Boosters Carried Maximum Payload

```
%%sql

Select BOOSTER_VERSION, PAYLOAD_MASS__KG_ from SPACEXDATASET
where PAYLOAD_MASS__KG_= (select MAX(PAYLOAD_MASS__KG_) from SPACEXDATASET);
```

```
 * ibm_db_sa://stf29843:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb
Done.
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

Booster versions which have carried the maximum payload mass

2015 Launch Records

The failed landing in drone ship in year 2015:

```
%%sql

Select "Landing _Outcome", BOOSTER_VERSION, LAUNCH_SITE, DATE from SPACEXDATASET
 where "Landing _Outcome" like '%Failure (drone ship)%'
    and Year(Date)=2015;
```

```
 * ibm_db_sa://stf29843:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90l08kqb1od81
Done.
```

Landing _Outcome	booster_version	launch_site	DATE
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40	2015-04-14

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

First stage landing outcomes between 2010-06-04 and 2017-03-20:

```
%%sql

Select "Landing _Outcome", Count(*) as Rank_Landing_Outcomes from SPACEXDATASET where DATE between '2010-06-04'and'2017-03-20'
group by "Landing _Outcome" order by Rank_Landing_Outcomes DESC;
```

```
 * ibm_db_sa://stf29843:***@824dfd4d-99de-440d-9991-629c01b3832d.bs2io90108kqb1od8lcg.databases.appdomain.cloud:30119/bludb
Done.
```

Landing _Outcome	rank_landing_outcomes
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

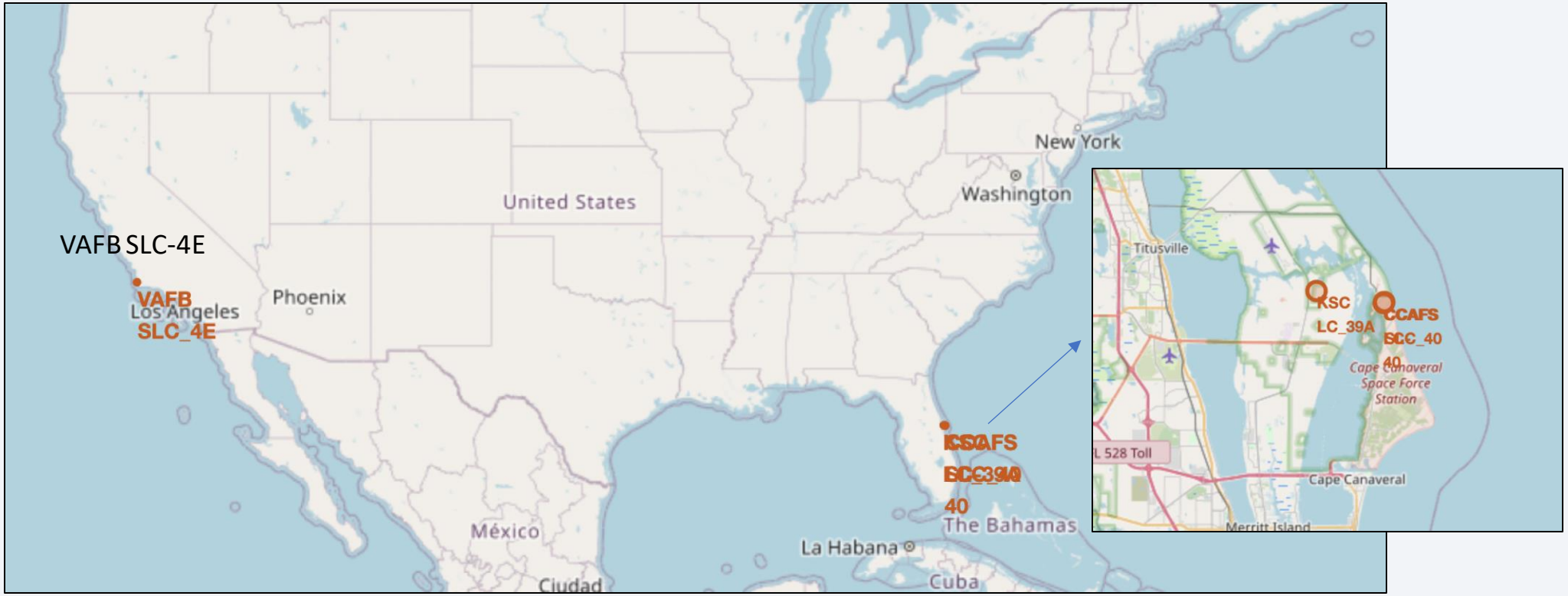
Insights from Interactive Visual Analytics

Section 3a

Launch Sites Proximities Analysis

Results – Visual Analytics with Folium

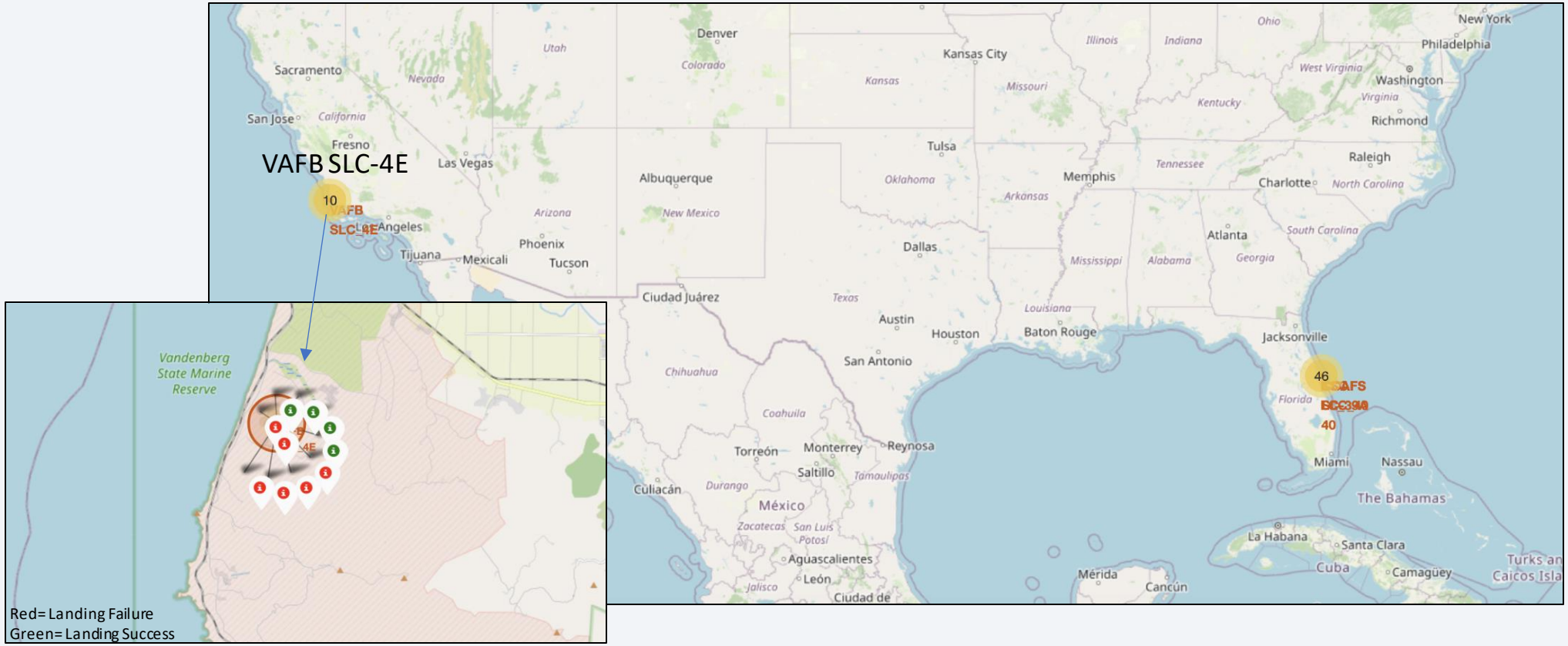
Launch Sites Location Map



- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A

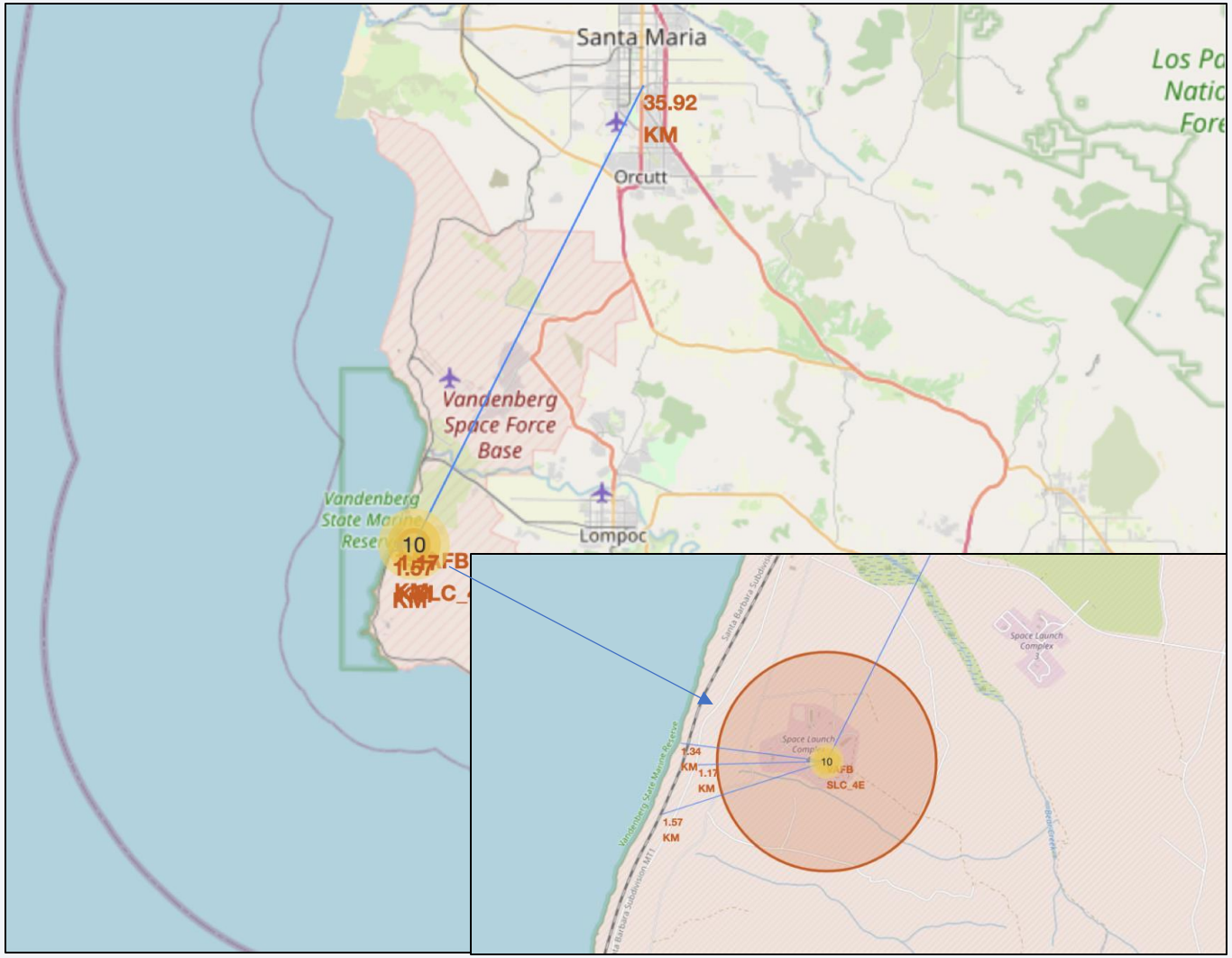
All launch sites are located in the US: 3 in Florida and 1 in California

Success and Failure Launches Map



- Total of 46 launches for the Florida sites
- Total of 10 launches for the California
- VAFB SLC-4E site: 4 launches were successful and 6 failed

Proximities Map – VAFB SLC-4E Site



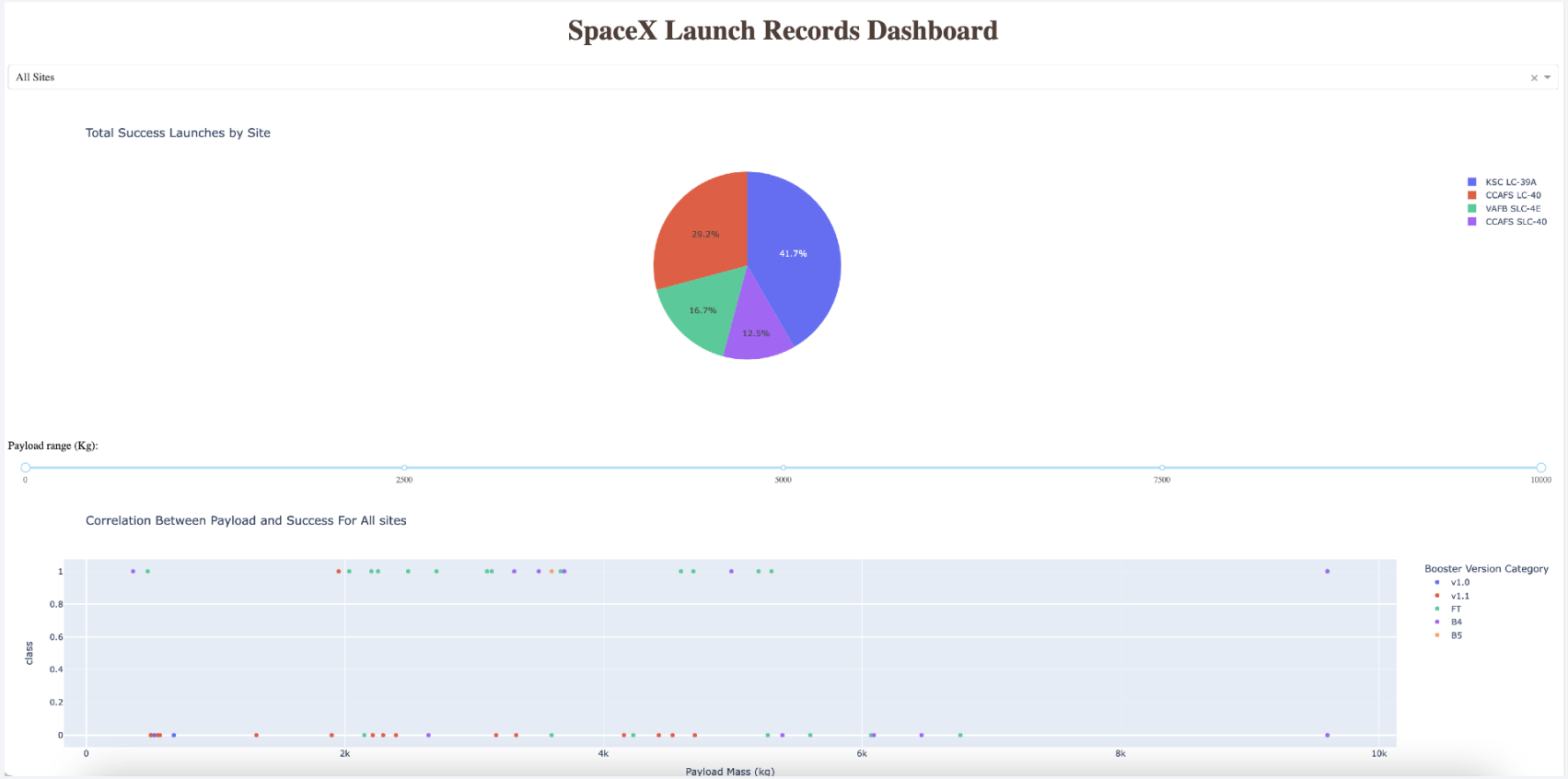
- VAFB SLC-4E is located close to the Pacific Ocean coast (1.34 km)
- VAFB SLC-4E is located close to infrastructures: the closest highway and railway are located 1.17km and 1.57km respectively from the site
- The closest city to VAFB SLC-4E site is Santa Maria (35.92km)

Section 3b

Build a Dashboard with Plotly Dash

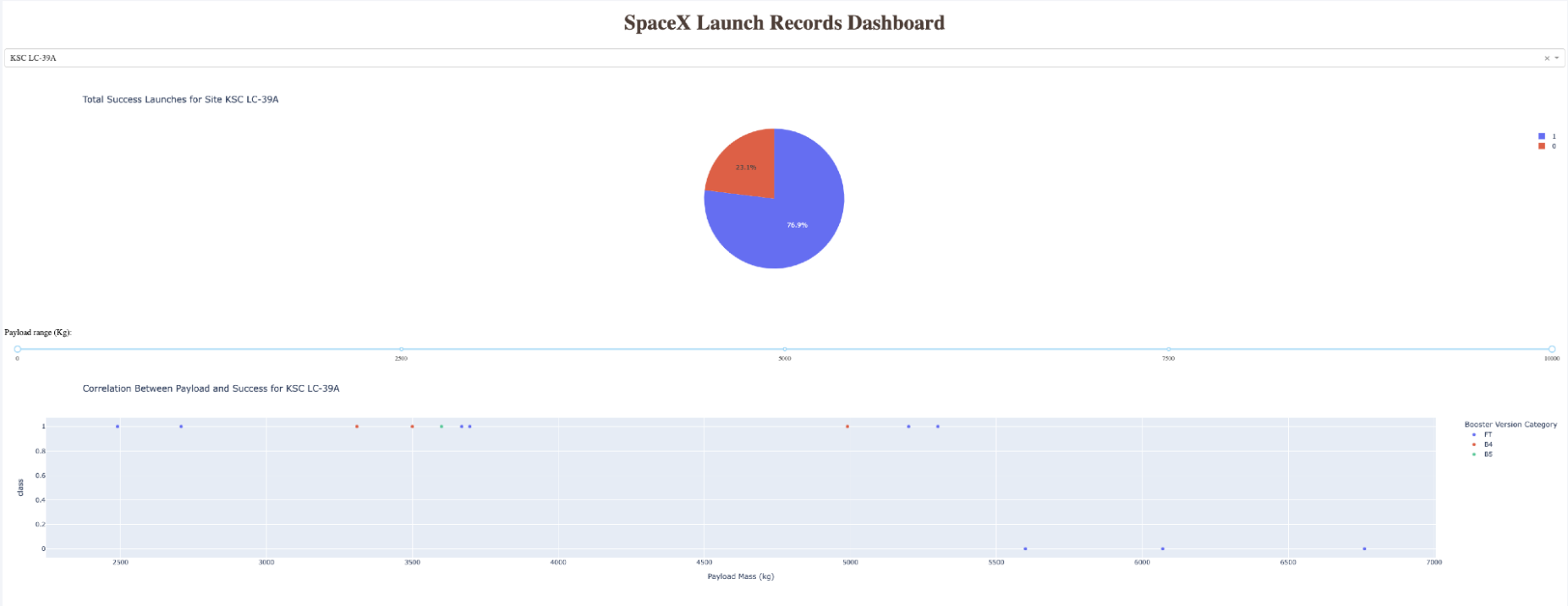
Results – Visual Analytics with Plotly Dash

Launch Records Dashboard – All Sites



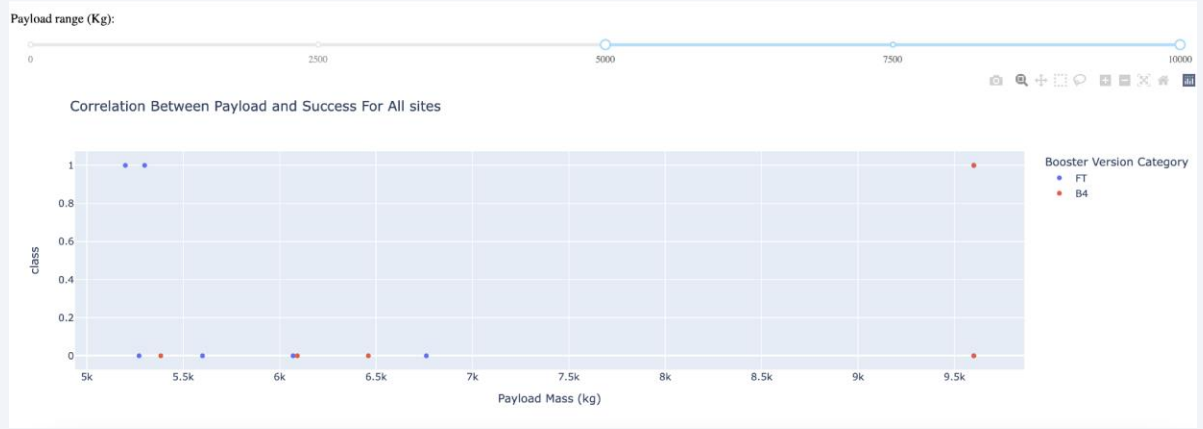
- From the pie chart above it is possible to observe that the site with the highest success is KSC LC-39A
- From the scatter plot below it is possible to see that the booster version FT is linked to the highest number of successful launches for all launch sites

Launch Records Dashboard – KSC LC-39A



- The success rate for KSC LC-39A site is 76.9% (failure rate 23.1%)
- Most of the launches at KSC LC-39A site were performed with FT booster version

Payload Mass and Success Correlation – All Sites



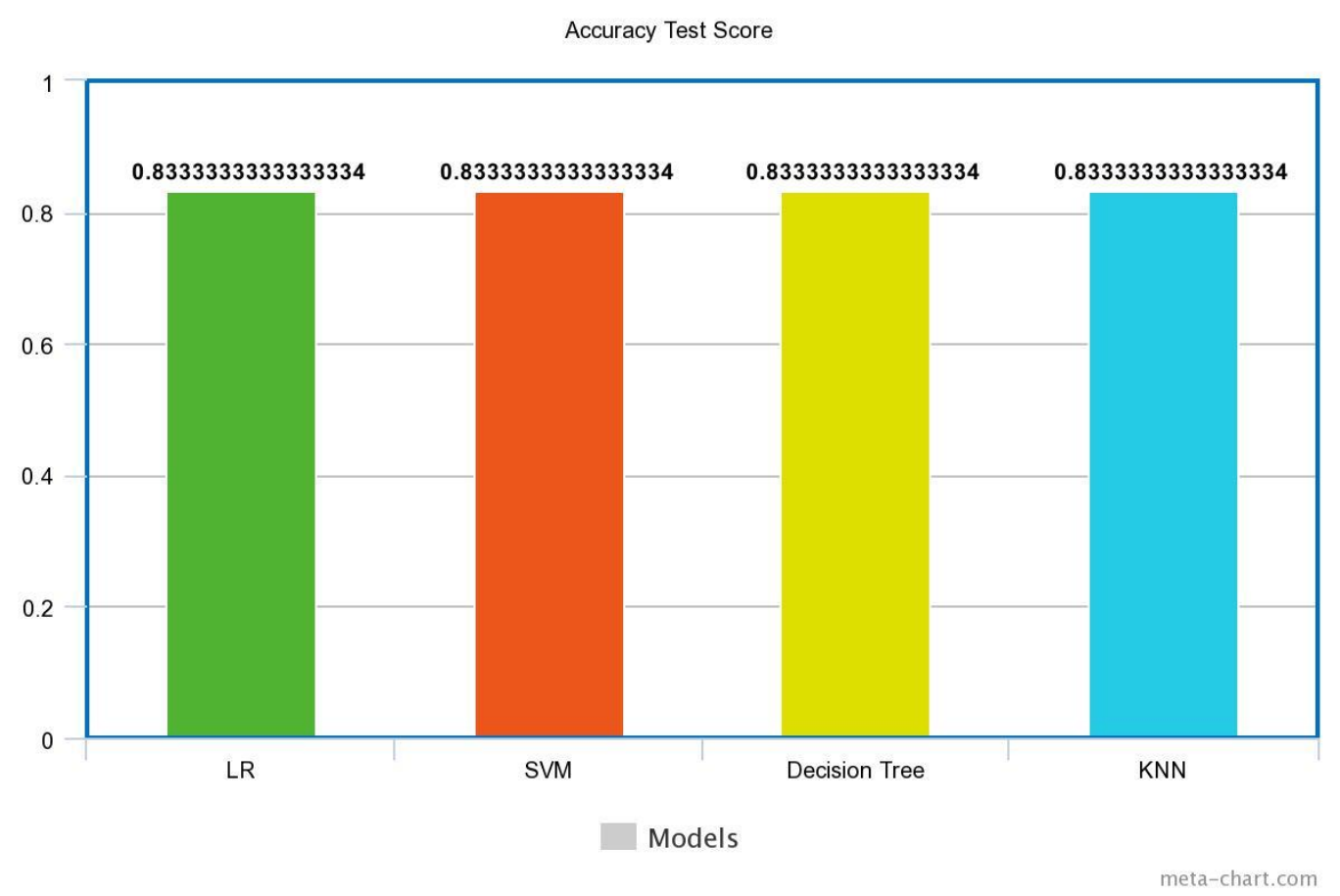
- The most successful booster versions are FT and B4 for any payload range (top scatter plot)
- The rate of launch success is very low for massive payload mass (more than 5000kg – right scatter plot) despite the booster version used for the launch
- For lower payload mass (less than 5000kg – left scatter plot) the most successful booster versions are FT and B4

Section 4

Predictive Analysis (Classification)

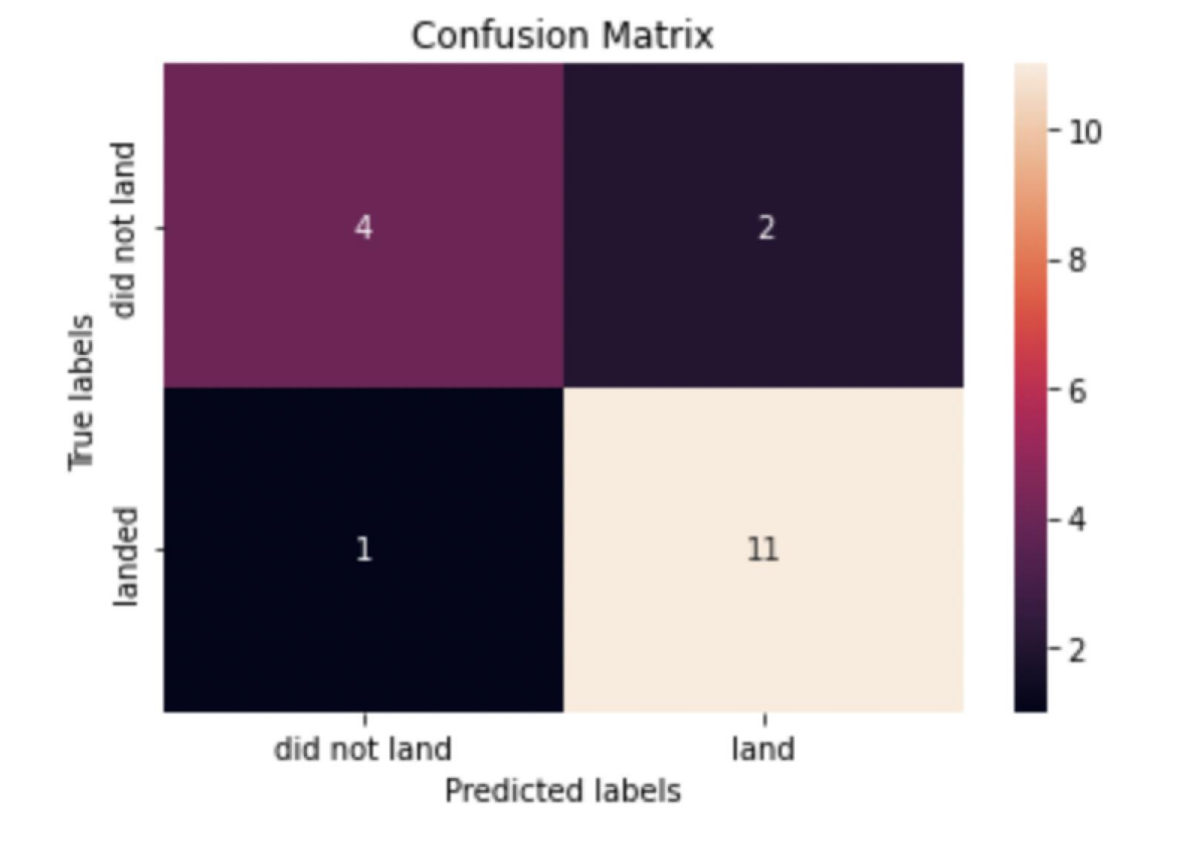
Results

Classification Accuracy – Test Set



- The accuracy score calculated on the test set is the same for all the 4 classification models
- The model with the highest accuracy score for the training set is the Decision Tree (~0.88%)

Confusion Matrix – Decision Tree



- 11 true positives (points predicted accurately - successful landing)
- 4 true negatives (points predicted accurately - failed landing)
- 2 false positives (launches predicted as successful when they were actually failures)
- 1 false negatives (launches predicted as failure when they were actually successful)

Conclusions

1. From the analysis undertaken it's possible to affirm that the factors having the highest impact on the successful landing of the Falcon 9 first stage are:

- Payload Mass Weight
- Booster Version
- Launch Site Location
- Number of continuous attempt launches
- Orbit Type

2. Classification models tested gave very similar performance results; the Decision Tree model seems slightly better in giving reliable predictions

3. Utilising the Decision Tree model and the knowledge gain from the analysis, it is possible to make predictions if the Falcon 9 first stage will land successfully for future launches. If the first stage land successfully a rocket launch will cost Space X about 62 million $ as advertised

Further Work Recommendations

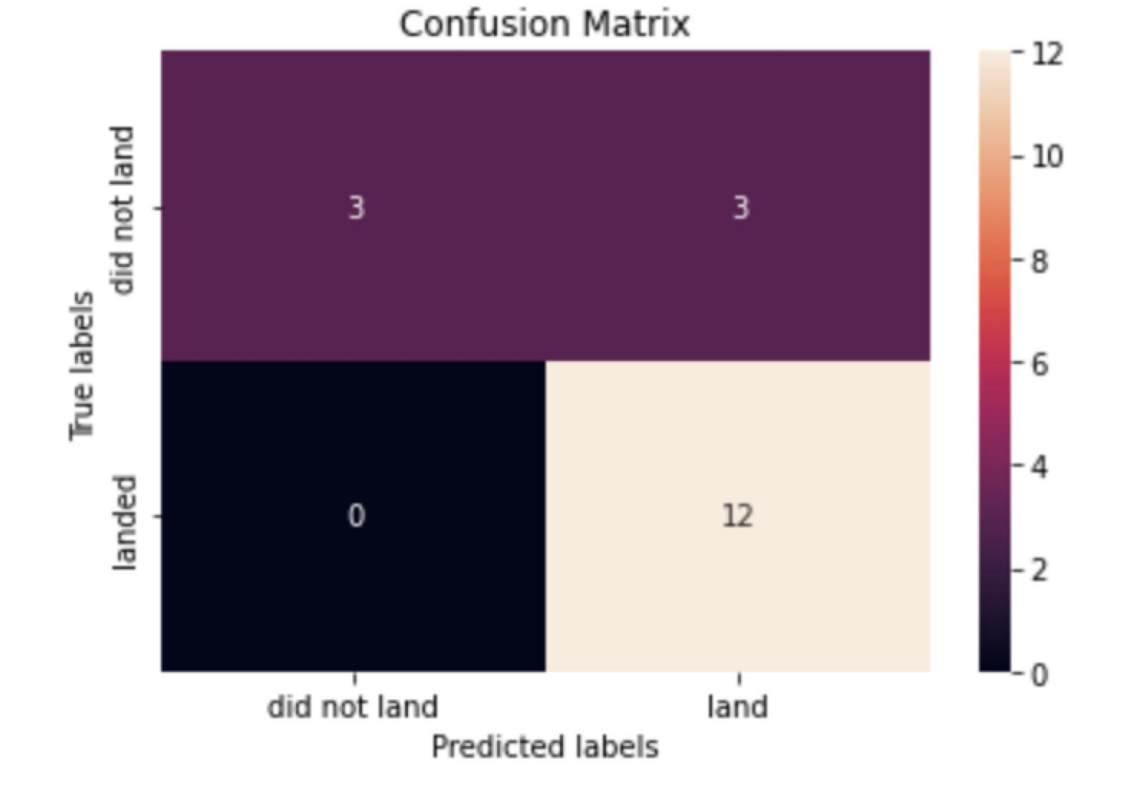
- Gather more data since the records obtain from SpaceX API and Wikipedia were limited:

i. Other parameters not highlighted in this project might result in having an effect on the successful landing of the Falcon 9 first stage

ii. A different classification model could be more suitable to make predictions compared to the one recommended in this project

- Do a more in-depth data analysis to get better insights about the relationship between parameters analyzed in this project
- Investigate if the landing locations (on drone ship, on ground etc.) has an effect on the landing outcome

Appendix – Confusion Matrix for other models



- 12 true positives (points predicted accurately - successful landing)
- 3 true negatives (points predicted accurately - failed landing)
- 3 false positives (launches predicted as successful when they were actually failures)
- 0 false negatives (launches predicted as failure when they were actually successful)

Thank you!