IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Vicente Espinosa
29-07-2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis (EDA)
  - Visualization
  - Predictive Models
- Summary of all results
  - Records collected from Falcon 9 landing outcomes
  - Four launch sites on south US which have fast access
  - Not clear relation between payload and other variables, except for flight numbers
  - High success rate of landing
  - Models predict with the same accuracy unknow data

Introduction

- Companies are making space travel affordable
- Space X has been the most successful, mainly due to cost reductions by recovering phase one stage
- The objective is to determine each price launch and if phase one will be reused due to a machine learning prediction on public information

Section 1

Methodology

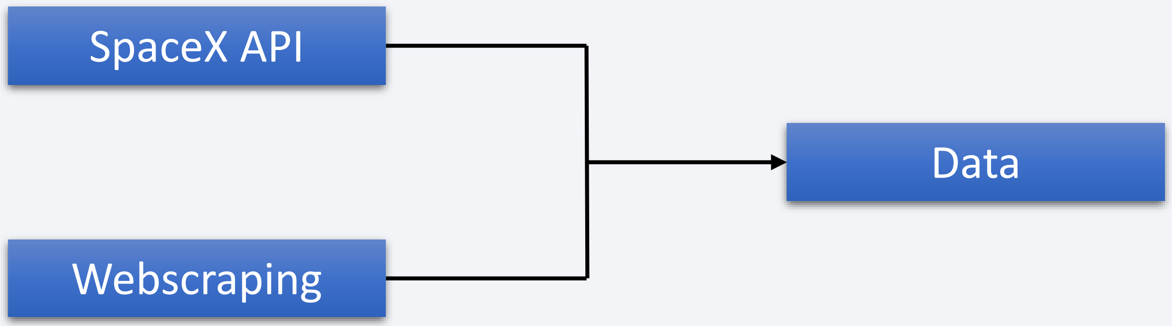
Methodology

Executive Summary

- Data collection methodology:
  - Data collected by SpaceX API and webscraping
- Perform data wrangling
  - Exploratory data analysis (EDA) with python and determine labels for training supervised model
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

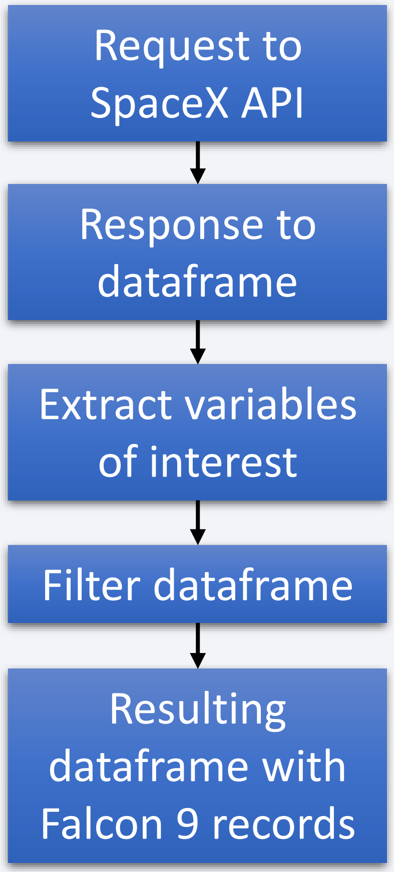
Data Collection

- The data was requested to the Space X API and then cleaned.
- Also, it was webscraped from Falcon 9 historical launch records from a Wikipedia page



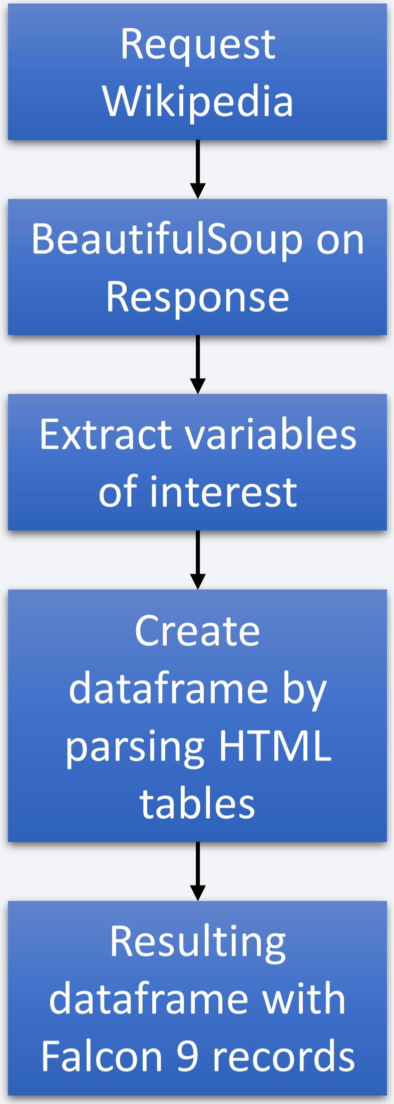
Data Collection - SpaceX API

- The data was requested to the Space X API
- Follow the next link to see the code used
- GitHub URL



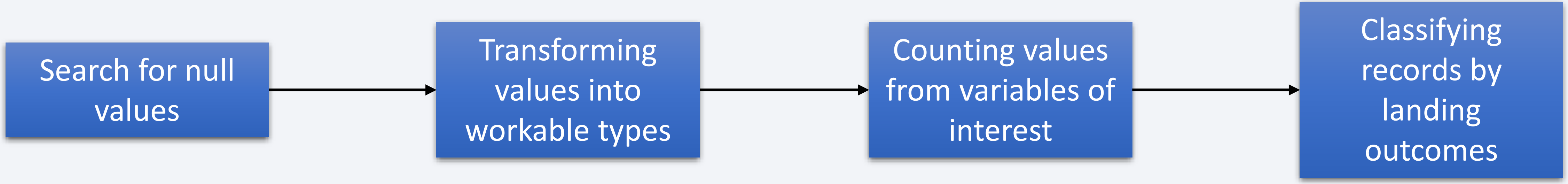
Data Collection - Scraping

- The data was webscraped from Falcon 9 historical launch records from a Wikipedia page
- Follow the next link to see the code used
- GitHub URL



Data Wrangling

- Exploratory Data Analysis (EDA) with python to find some patterns in the data and determine what would be the label for training supervised models
- GitHub URL



EDA with Data Visualization

- Mainly scatter plots to see the different relations between the variables of interest.
- A bar chart to see the relation between success and orbit type.
- A line plot to analyze the success yearly trend
- GitHub URL

EDA with SQL

- Determine unique launch sites
- Determine total payload mass per client and avg payload mass per booster version
- Find date, booster version, payload mass and type of success and failure records between dates and/or booster version
- GitHub URL

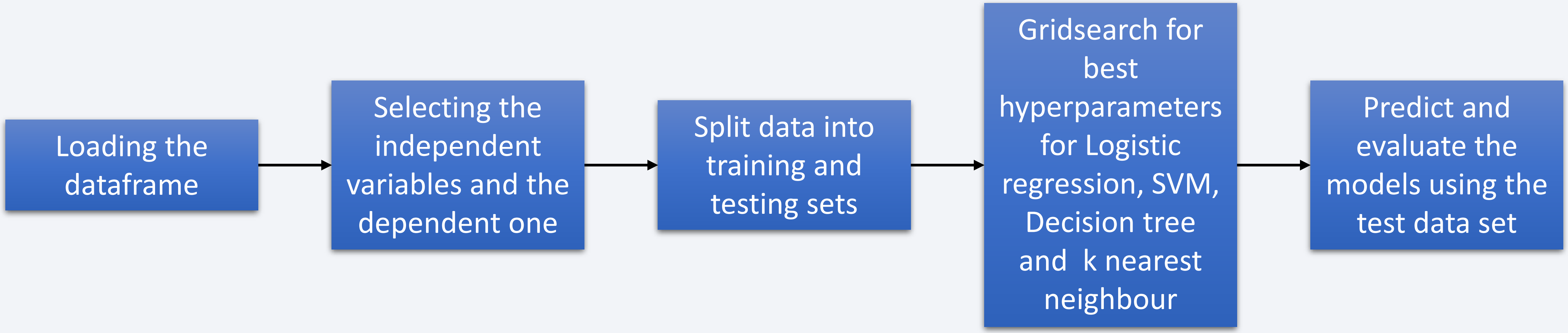
Build an Interactive Map with Folium

- The map has markers, circles, clusters and lines
    - Markers shows landing outcome, classified by success (green) and failure (red)
    - Circles show the launch site
    - Clusters are made by the markers inside the launch site circles
    - Lines show distances between a launch site and nearest coastline, railway, highway and city
    - GitHub URL

Build a Dashboard with Plotly Dash

- Dashboard contains
    - Dropdown of the launch sites to select between all or each site, modifying in real time the graphs
    - Pie chart showing success failure ratio for selected sites
    - Range slider to select payload to limit the range of the payload mass showed in the scatter plot
    - Scatter plot between payload mass and mission outcome for selected site to analyze possible correlations between payload and mission outcome.
- GitHub URL

Predictive Analysis (Classification)



- GitHub URL

Results

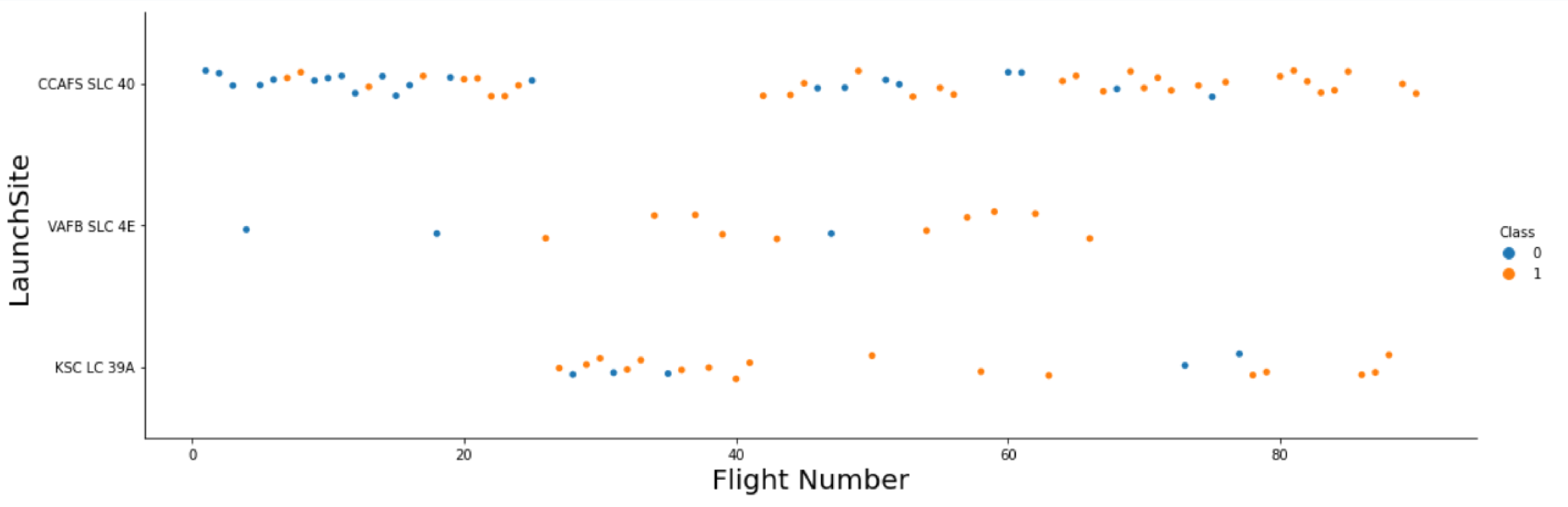
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

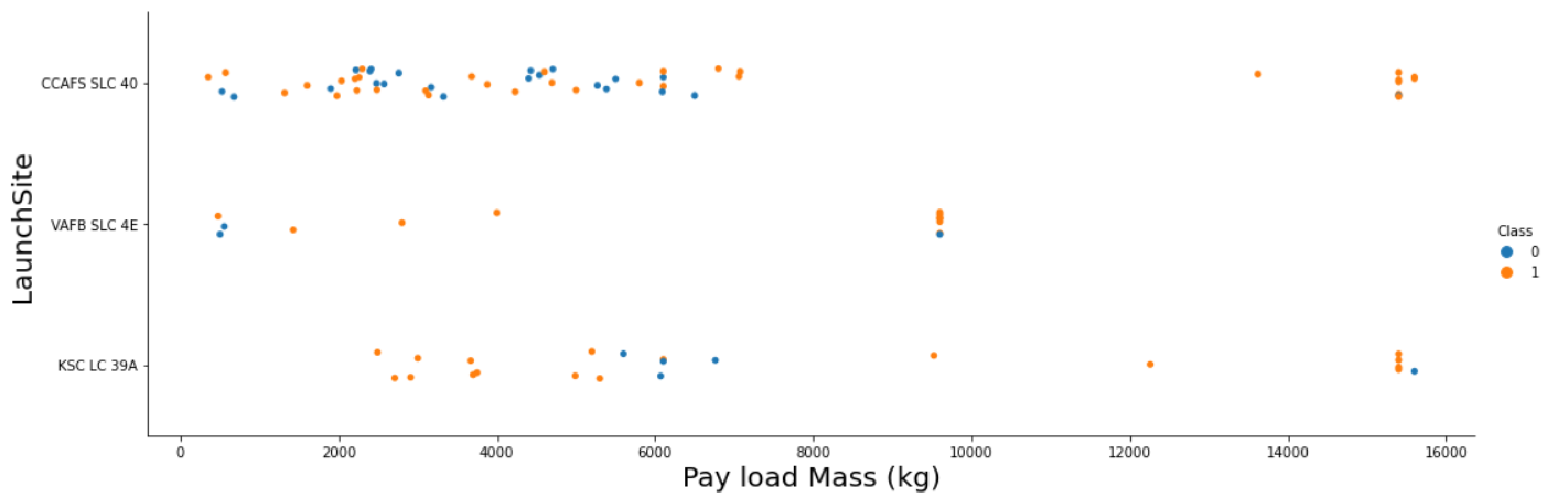
Flight Number vs. Launch Site

- Relation between Flight Number success (Class 1) per Launch Site



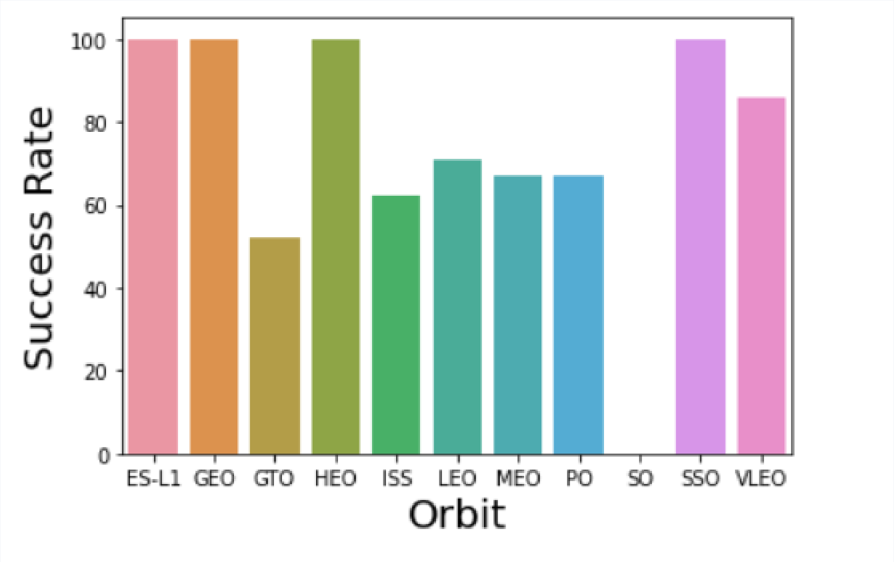
Payload vs. Launch Site

- Relation between payload mass success (Class 1) per Launch Site



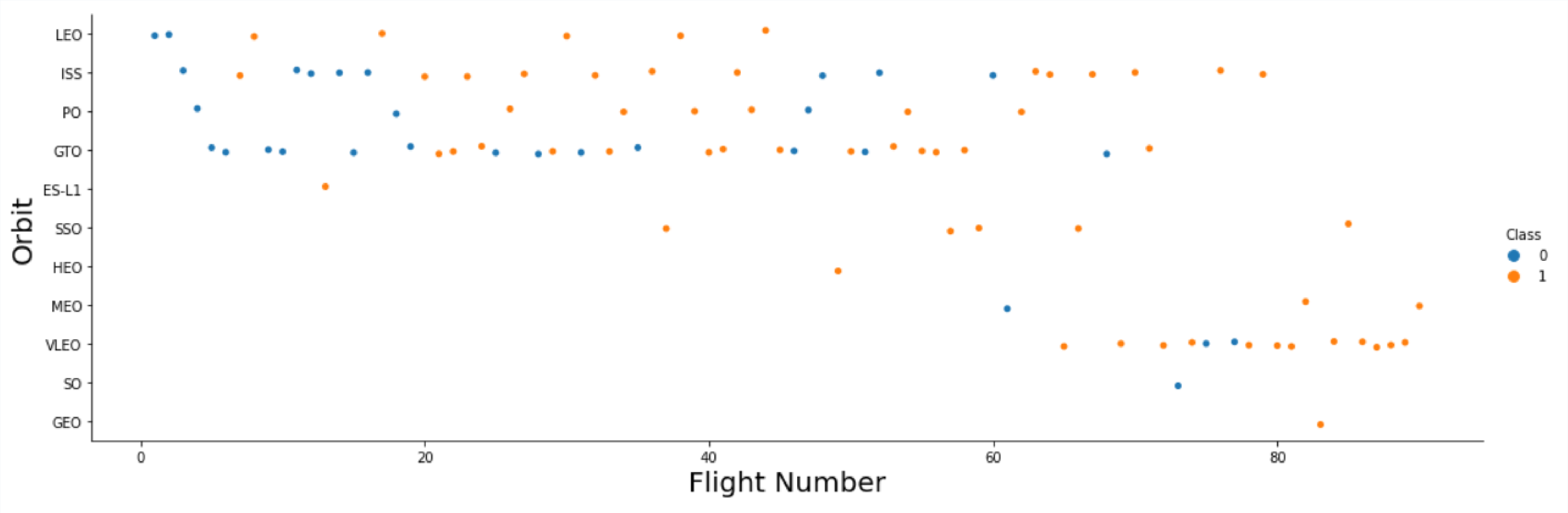
Success Rate vs. Orbit Type

- Bar chart showing the success rate per orbit



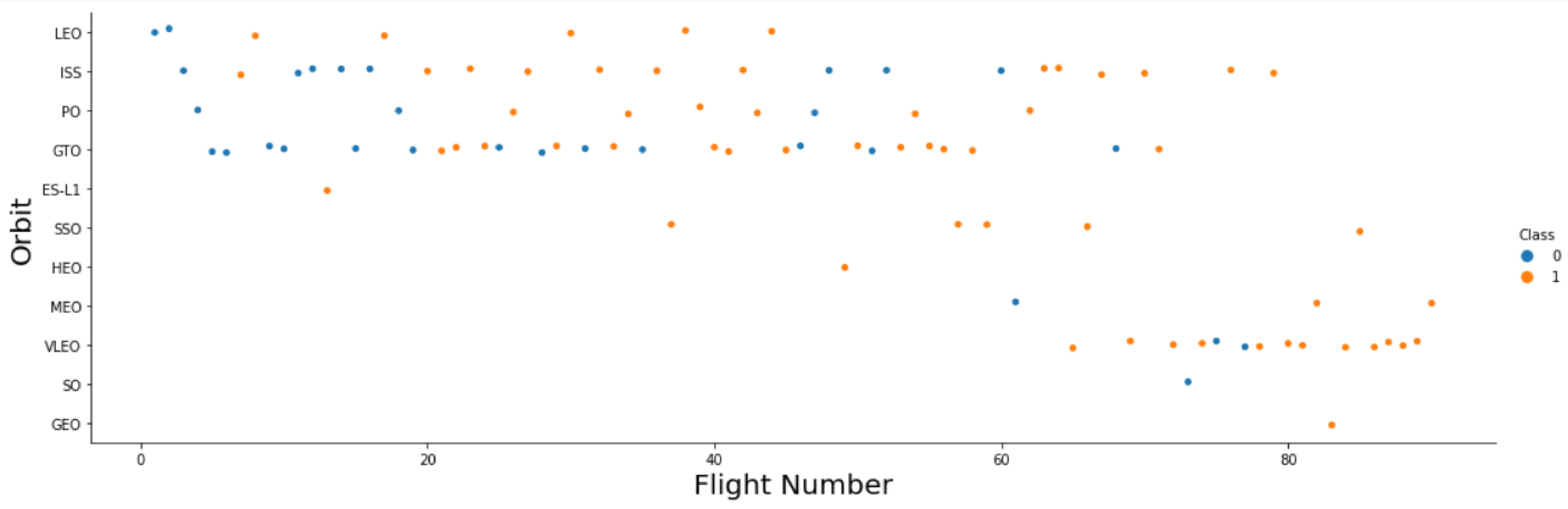
Flight Number vs. Orbit Type

- Relation between Flight Number success (Class 1) per orbit



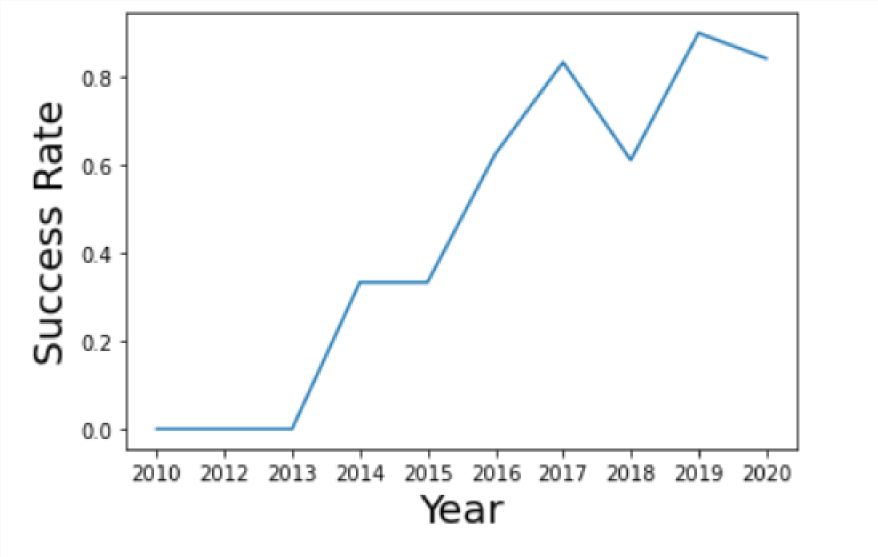
Payload vs. Orbit Type

- Relation between Flight Number success (Class 1) per orbit



Launch Success Yearly Trend

- Line chart of yearly average success rate showing the general trend
- Upward trend since 2013



All Launch Site Names

- Names of the unique launch sites
- We have four launch sites

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA` alluding to Cabe Canaveral launch site

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	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

Customer	Total_Payload_Mass_NASA
NASA (CRS)	45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1

Booster_Version	Avg_Payload_Mass_F9v11
F9 v1.1 B1003	2534.6666666666665

First Successful Ground Landing Date

- First successful landing outcome on ground pad date

min(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

Booster_Version	Landing _Outcome	PAYLOAD_MASS__KG_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes
- Showing a great number of successes over failures

Landing _Outcome	Number_Outcomes
Success	38
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Failure	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1
No attempt	1

Boosters Carried Maximum Payload

- Booster which have carried the maximum payload mass
- Corresponding to B5 version

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

2015 Launch Records

- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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

- Landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

Landing _Outcome	Number_Outcomes
Success	20
Success (drone ship)	8
Success (ground pad)	6

Section 3

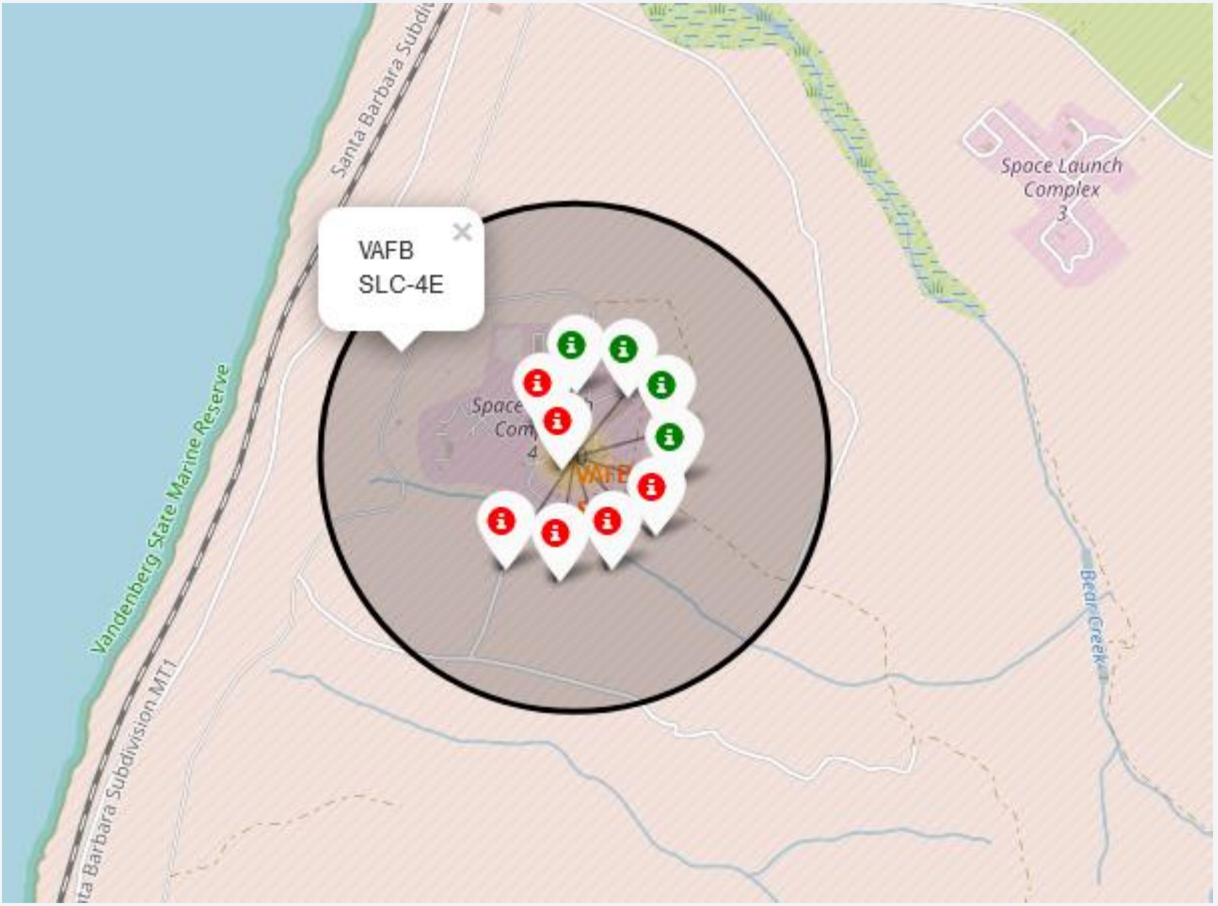
Launch Sites Proximities Analysis

Launching Sites

- All launch sites located at south states of the US
- Launches in California and Florida
- Mainly from Florida with 82% of launches

VAFB SLC-4E

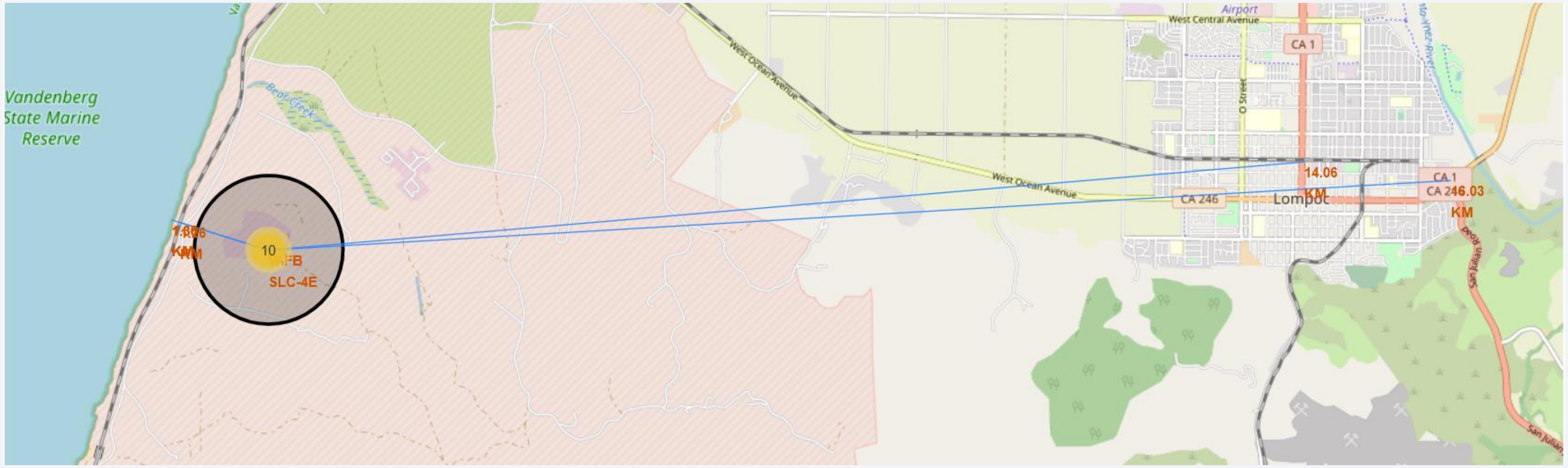
- Launches from Vandenberg Space Launch Complex, California
- Classified by landing outcomes, where success is green and failure red
- 40% of success



VAFB SLC-4E Proximities

- Proximities to Vandenberg Space Launch Complex
- Distances shown in the table

	Proximity	Distance [km]
0	City	14.064943
1	Railway	1.255995
2	Highway	16.029340
3	Coastline	1.363247

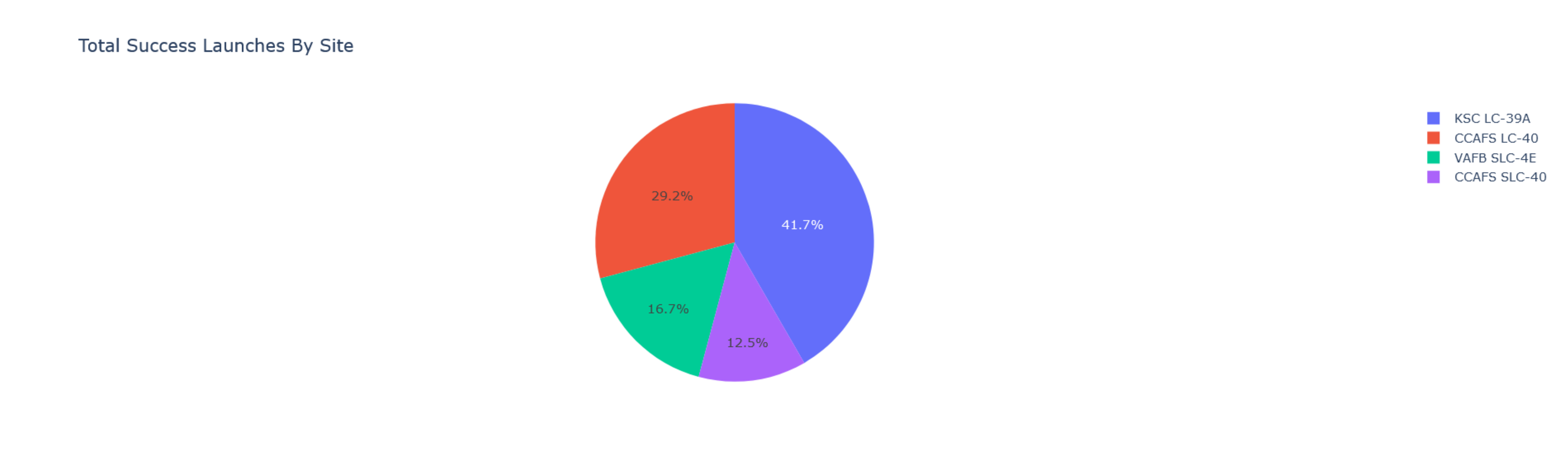


Section 4

Build a Dashboard with Plotly Dash

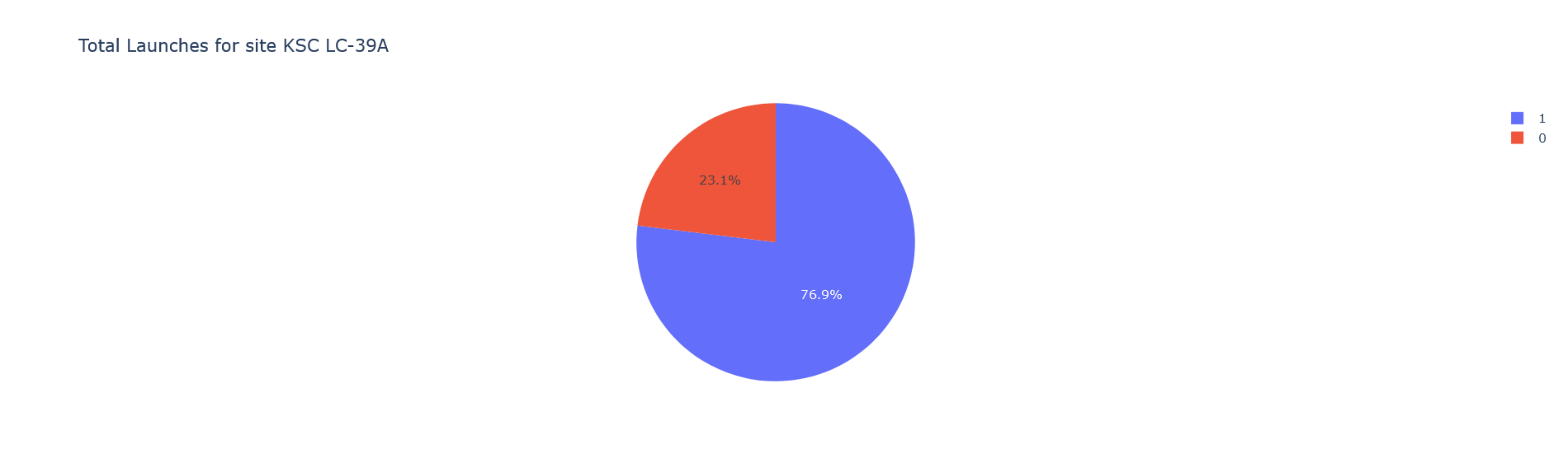
Success Launches by Site Pie Chart

- Success count of all launch sites
- Better success ratio from KSC LC39A
- Worst success ratio from CCAFS SLC-40



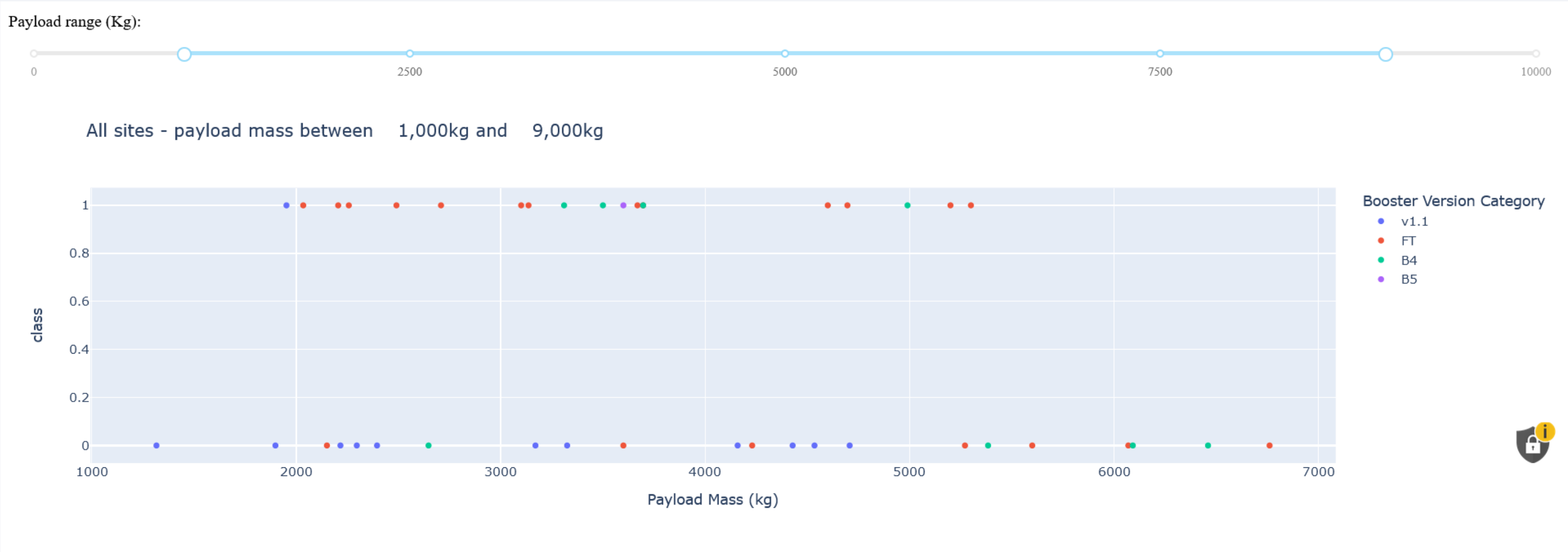
Success Launches from KSC LC-39A

- Pie chart of the best success ratio launch site
- Proving a near 77% successful landing outcomes



Payload vs. Launch Outcomes Scatter Plot

- Scatter plot doesn't show a clear relation between payload, booster version and success rate.

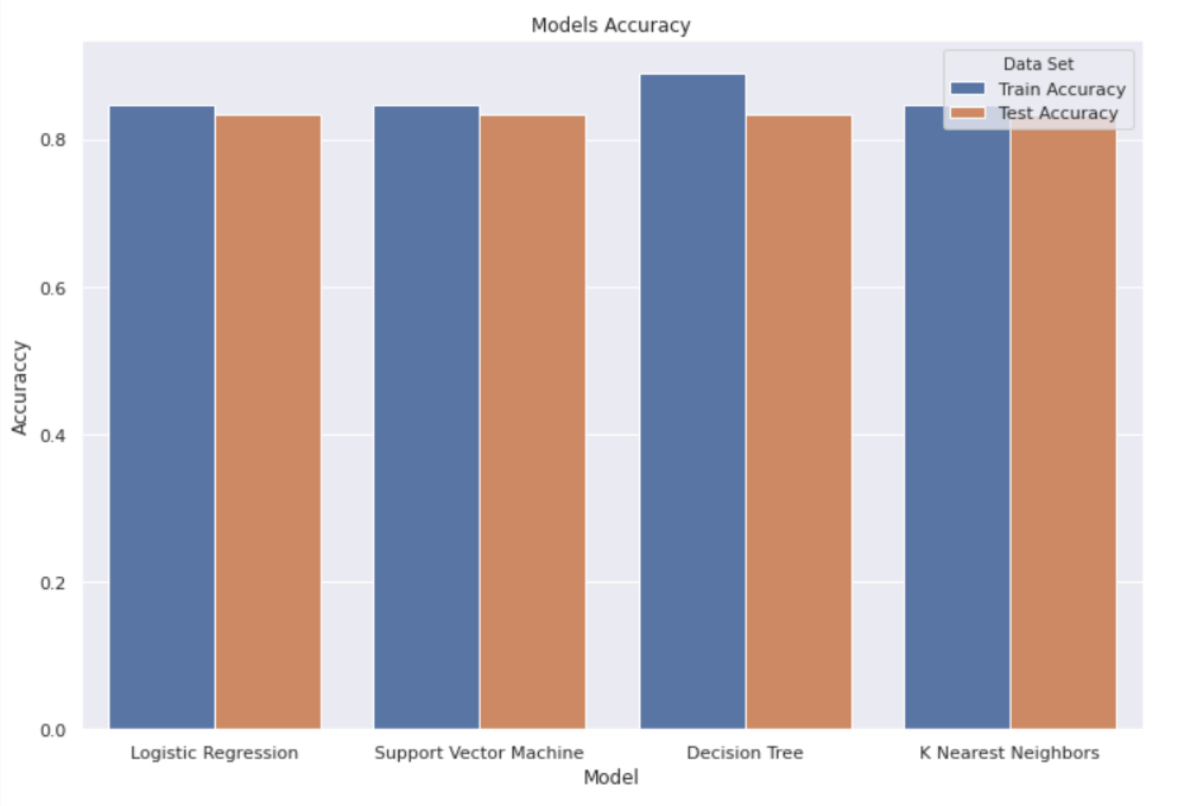


Section 5

Predictive Analysis (Classification)

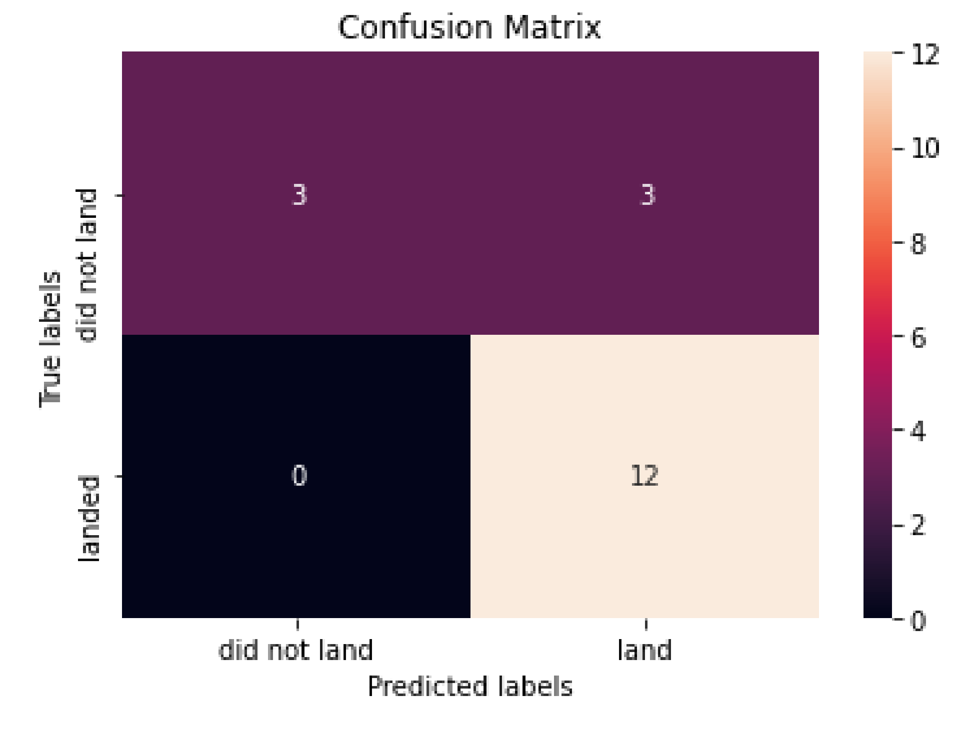
Classification Accuracy

- All the models present the same train accuracy, except for the decision tree which is a little bit higher
- The four models presents the same test accuracy
- The models will perform the same with the unknow data, except for decision tree which could capture noise



Confusion Matrix

- All the models show the same confusion matrix
- This confusion matrix show a good accuracy for successful landing outcome
- Also, show accuracy of 50% in failures in lading outcomes



Conclusions

- Only four launching sites
- High overall success in landing outcomes
- Landing outcomes have a successful upward trend since 2013
- Not clear correlation between orbit and successful landing, same for payload, booster version and success rate
- Launch sites on south of the US (Florida and California), nearest the Ecuador
- Launch sites have proximities with coastlines, railways, highways and a city
- Most successful landing outcome rate is from KSC LC39A and worst from CCAFS SLC-40
- The predictive models perform the same with the unknow data, except for decision tree which could capture noise
- All the models show the same confusion matrix which has a 100% accuracy for successful landing outcomes and 50% for failure ones.
- This prove that phase one have a high successful rate to be recovered

Appendix

- Code for proximities table

```
launch_site_lat = 34.632834
launch_site_lon = -120.610745
loc = ['City', 'Railway', 'Highway', 'Coastline']
dic = {loc[0]:[34.64347, -120.4576],loc[1]: [34.63616, -120.62386],loc[2]: [34.64123, -120.43589], loc[3]: [34.63639, -120.625]}
dist = []
for i in loc:
    lat = dic[i][0]
    long = dic[i][1]
    distance = calculate_distance(launch_site_lat, launch_site_lon, lat, long)
    dist.append(distance)
d = pd.DataFrame({'Proximity': loc, 'Distance [km]': dist})
d
```

Appendix

- Code for model accuracy for all built classification models

```
model = ['Logistic Regression', 'Support Vector Machine', 'Decision Tree', 'K Nearest Neighbors']
models = model*2
train = [logreg_cv.best_score_, svm_cv.best_score_, tree_cv.best_score_, knn_cv.best_score_]
test = [logreg_cv.score(X_test, Y_test), svm_cv.score(X_test, Y_test), tree_cv.score(X_test, Y_test), knn_cv.score(X_test, Y_test)]
dt = train + test
clss = ['Train Accuracy']*4 + ['Test Accuracy']*4
dfm = pd.DataFrame({'Model':models, 'Accuracy':dt, 'Data Set':clss})
sns.set(rc={'figure.figsize':(12,8)})
sns.barplot(x='Model', y='Accuracy', hue='Data Set', data=dfm)
plt.title('Models Accuracy')
plt.xlabel('Model')
plt.ylabel('Accuraccy')
plt.show()
```

Thank you!