IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Jason Forsythe
7/14/22

Outline

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

Executive Summary

- Summary of methodologies
- Summary of all results

Introduction

In this project I wanted to find out the answers to a few questions

- Can I predict the success of a SpaceX first stage landing?
- If so, what parameters could be used to give the best prediction?
- What would be the best accuracy of prediction that I could achieve?

Section 1

Methodology

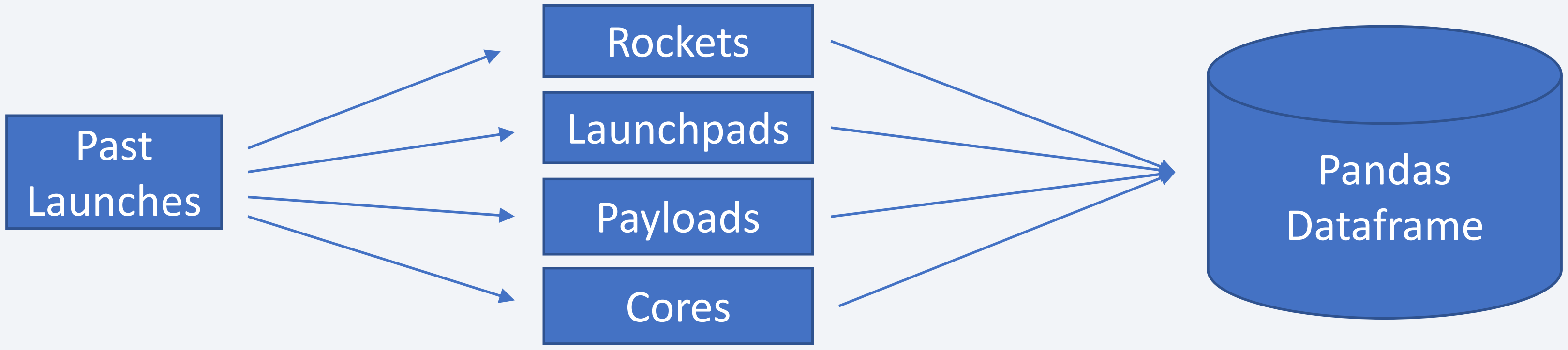
Methodology

Executive Summary

- Data collection methodology:
  - Data was collected via SpaceX API and Web Scraping https://github.com/Jason-AE/IBM_Data_Science/tree/master/DataCollectionAPI
- Perform data wrangling
  - Once Data was collected, I removed unesseary columns and replaced null values. https://github.com/Jason-AE/IBM_Data_Science/tree/master/DataWrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - I tested Logistic Regression, Vector Machine, Tree, and K Nearest Neighbor models using grid search to find the best parameters for each. I then compared the results using test data and the Tree model performed the best. https://github.com/Jason-AE/IBM_Data_Science/tree/master/PredictiveAnalysis

Data Collection – SpaceX API

- Data was collected using SpaceX's own Web API
  - https://api.spacexdata.com/v4/launches/past
  - This data was used along with different endpoints to combine data about the rockets, launchpads, payloads, and cores.
  - The Data was then stored in a Panda's Data Frame



https://github.com/Jason-AE/IBM_Data_Science/tree/master/DataCollectionAPI

Data Collection - Scraping

- Data was also collected using Web scrapping using Wikipedia data in HTML Tables
  https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

[hide] Flight No.	Date and time (UTC)	Version, Booster [b]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
1	4 June 2010, 18:45	F9 v1.0[7] B0003.1[8]	CCAFS, SLC-40	Dragon Spacecraft Qualification Unit		LEO	SpaceX	Success	Failure[9][10] (parachute)
	First flight of Falcon 9 v1.0.[11] Used a boilerplate version of Dragon capsule which was not designed to separate from the second stage.(more details below) Attempted to recover the first stage by parachuting it into the ocean, but it burned up on reentry, before the parachutes even deployed.[12]								
2	8 December 2010, 15:43[13]	F9 v1.0[7] B0004.1[8]	CCAFS, SLC-40	Dragon demo flight C1 (Dragon C101)		LEO (ISS)	NASA (COTS) NRO	Success[9]	Failure[9][14] (parachute)
	Maiden flight of Dragon capsule, consisting of over 3 hours of testing thruster maneuvering and reentry.[15] Attempted to recover the first stage by parachuting it into the ocean, but it disintegrated upon reentry, before the parachutes were deployed.[12] (more details below) It also included two CubeSats,[16] and a wheel of Brouère cheese.								
3	22 May 2012, 07:44[17]	F9 v1.0[7] B0005.1[8]	CCAFS, SLC-40	Dragon demo flight C2+[18] (Dragon C102)	525 kg (1,157 lb)[19]	LEO (ISS)	NASA (COTS)	Success[20]	No attempt
	Dragon spacecraft demonstrated a series of tests before it was allowed to approach the International Space Station. Two days later, it became the first commercial spacecraft to board the ISS.[17] (more details below)								
4	8 October 2012, 00:35[21]	F9 v1.0[7] B0006.1[8]	CCAFS, SLC-40	SpaceX CRS-1[22] (Dragon C103)	4,700 kg (10,400 lb)	LEO (ISS)	NASA (CRS)	Success	No attempt
				Orbcomm-OG2[23]	172 kg (379 lb)[24]	LEO	Orbcomm	Partial failure[25]	
	CRS-1 was successful, but the secondary payload was inserted into an abnormally low orbit and subsequently lost. This was due to one of the nine Merlin engines shutting down during the launch, and NASA declining a second reignition, as per ISS visiting vehicle safety rules, the primary payload owner is contractually allowed to decline a second reignition. NASA stated that this was because SpaceX could not guarantee a high enough likelihood of the second stage completing the second burn successfully which was required to avoid any risk of secondary payload's collision with the ISS.[26][27][28]								
5	1 March 2013, 15:10	F9 v1.0[7] B0007.1[8]	CCAFS, SLC-40	SpaceX CRS-2[22] (Dragon C104)	4,877 kg (10,752 lb)	LEO (ISS)	NASA (CRS)	Success	No attempt
	Last launch of the original Falcon 9 v1.0 launch vehicle, first use of the unpressurized trunk section of Dragon.[29]								
6	29 September 2013, 16:00[30]	F9 v1.1[7] B1003[8]	VAFB, SLC-4E	CASSIOPE[22][31]	500 kg (1,100 lb)	Polar orbit LEO	MDA	Success[30]	Uncontrolled (ocean)[d]
	First commercial mission with a private customer, first launch from Vandenberg, and demonstration flight of Falcon 9 v1.1 with an improved 13-tonne to LEO capacity.[29] After separation from the second stage carrying Canadian commercial and scientific satellites, the first stage booster performed a controlled reentry,[32] and an ocean touchdown test for the first time. This provided good test data, even though the booster started rolling as it neared the ocean, leading to the shutdown of the central engine as the roll depleted it of fuel, resulting in a hard impact with the ocean.[30] This was the first known attempt of a rocket engine being lit to perform a supersonic retro propulsion, and allowed SpaceX to enter a public-private partnership with NASA and its Mars entry, descent, and landing technologies research projects.[33] (more details below)								
7	3 December 2013, 22:41[34]	F9 v1.1 B1004	CCAFS, SLC-40	SES-8[22][35][36]	3,170 kg (6,990 lb)	GTO	SES	Success[37]	No attempt [38]
	First Geostationary transfer orbit (GTO) launch for Falcon 9,[35] and first successful reignition of the second stage.[39] SES-8 was inserted into a Super-Synchronous Transfer Orbit of 79,341 km (49,300 mi) in apogee with an inclination of 20.55° to the equator.								

HTML Table Example

- BeautifulSoup Library Was Used to Extract Data
- Helper functions cleaned the HTML
- Pandas Dataframe was created
- Iterations over the rows was used to fill Dataframe

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	[[SpaceX], \n]	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	[[\n, [<a href="/wiki/NASA" title="NASA">N...	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	[[NASA], (, [COTS],)\n]	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	[[NASA], (, [CRS],)\n]	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	[[NASA], (, [CRS],)\n]	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10
...	...	...	...	...	...	...	...	...	...	...	...
116	117	CCSFS	Starlink	15,600 kg	LEO	[[SpaceX], \n]	Success\n	F9 B5B1051.10	Success	9 May 2021	06:42
117	118	KSC	Starlink	~14,000 kg	LEO	[[SpaceX], [], , [Capella Space], and , [Tyv...	Success\n	F9 B5B1058.8	Success	15 May 2021	22:56
118	119	CCSFS	Starlink	15,600 kg	LEO	[[SpaceX], \n]	Success\n	F9 B5B1063.2	Success	26 May 2021	18:59
119	120	KSC	SpaceX CRS-22	3,328 kg	LEO	[[NASA], (, [CRS],)\n]	Success\n	F9 B5B1067.1	Success	3 June 2021	17:29
120	121	CCSFS	SXM-8	7,000 kg	GTO	[[Sirius XM], \n]	Success\n	F9 B5	Success	6 June 2021	04:26

121 rows × 11 columns

Resulting Pandas Dataframe

https://github.com/Jason-AE/IBM_Data_Science/tree/master/DataCollectionAPI

Data Wrangling

- Once data was collected it needs to be cleaned
- Check for missing values in each attribute (~40 in LandingPad)
    - Replace missing values with mean in that column or remove rows
- Convert landing outcomes to a class of either 0 failed 1 success
- Now we can check the mean of success (66.666%)
- This clean up is important for normalization and predictive analytics (to be preformed later)

https://github.com/Jason-AE/IBM_Data_Science/tree/master/DataWrangling

EDA with Data Visualization

- In Exploratory Data Analysis I plotted the following charts:
- Success rate vs launch site: to determine if launch site affects success
- Flight number per site vs success rate: to determine if launches were more successful with later launches.
- Payload per site vs success rate: to determine if payload affects success rate at each site
- Success rate vs orbit type: to determine if orbit affects success rate
- Payload by orbit type vs success rate: to determine if certain payloads in a give orbit affect success rate
- Success rate over time (yearly): to determine if success increased over time
- https://github.com/Jason-AE/IBM_Data_Science/tree/master/EDA

EDA with SQL

- In exploratory data analysis using SQL I performed the following queries:
- Select distinct launch site names
- Displayed 5 records from the data set where the launch site contained 'KSC'
- Displayed the total payload carried for NASA missions
- Displayed the average payload carried by booster version F9 v1.1
- Showed the date of the first successful landing on a drone ship
- Listed the names of the boosters that have had successful landings on land with a mass greater than 4000 and less that 6000 kg
- Listed the total number of successful and failure mission outcomes
- Listed the booster versions that have carried the maximum payload mass
- For 2017 listed the booster versions and launch sites for successful landings on land
- Ranked the count of successful landings between 2010-06-04 and 2017-03-20 in descending order
- https://github.com/Jason-AE/IBM_Data_Science/tree/master/EDA

Build an Interactive Map with Folium

- I built a Folium Interactive Map to show the relation of launch sites and successful landings at each in a more visual way
- I used a folium.Circle object to mark the area of the launch site with details
- I also used a folium.map.Marker to mark a launch site when zoomed out
- I made markers of green and red to represent successful or failed launches
- Lastly, I made marker clusters to group successful and failed launches based on launch sites
- https://github.com/Jason-AE/IBM_Data_Science/tree/master/FoliumMap

Build a Dashboard with Plotly Dash

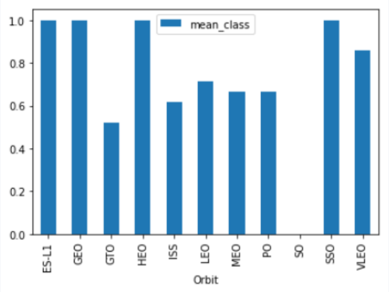
- I created an interactive dashboard with plotly with the following features:
- Pie chart showing success rate per launch site
- Scatter plot showing launch outcomes vs payload colored by booster version
- Launch site dropdown that includes all sites (Controls data for pie and scatter chart)
- Payload range slider (controls data for scatter chart only)
- https://github.com/Jason-AE/IBM_Data_Science/tree/master/Capstone_Mod3_Interactive_Dashboard

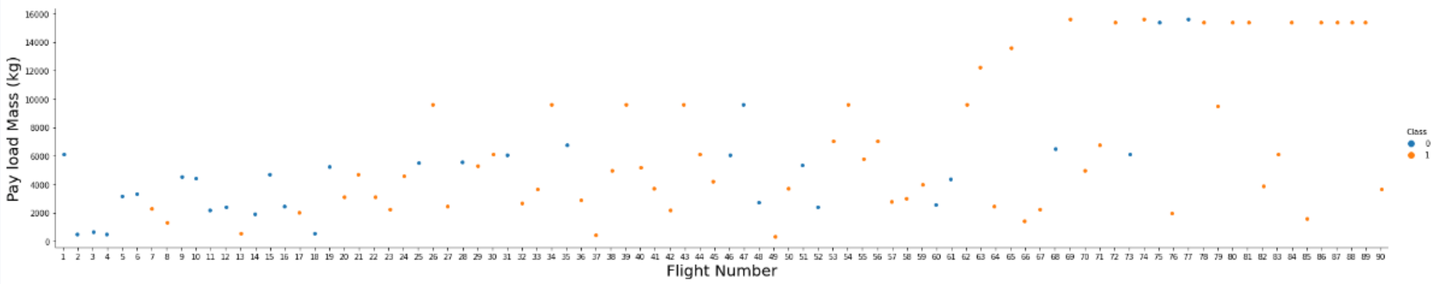
Predictive Analysis (Classification)

- During predictive Analysis I test the following models:
- Logistic Regression
- Vector Machine
- Tree
- K Nearest Neighbor
- I used grid search to find the best parameters for each.
- I then compared the results using test data and the Tree model performed the best.
  https://github.com/Jason-AE/IBM_Data_Science/tree/master/PredictiveAnalysis

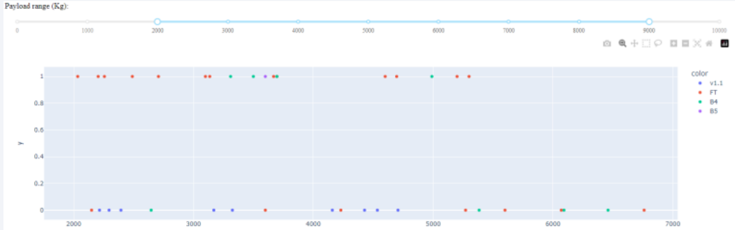
Results

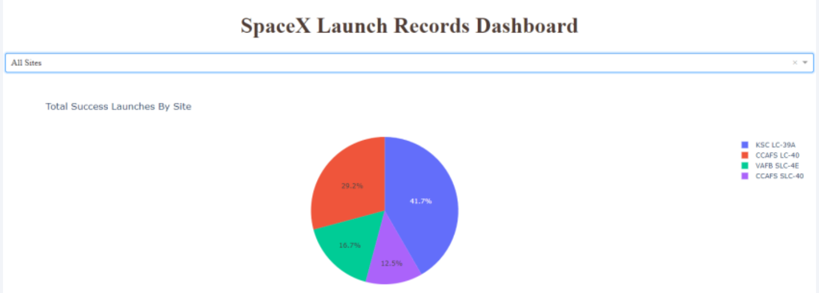
- I found that Orbit, Launch Site, Payload, and Flight Number are good predictors of success.





- Interactive analytics demo in screenshots





- Predictive analysis results

```
print("LogReg Score: ",logreg_cv.score(X_test, Y_test))
print("SVM Score: ", svm_cv.score(X_test, Y_test))
print("Tree Score: ", tree_cv.score(X_test, Y_test))
print("KNN Score: ", knn_cv.score(X_test, Y_test))
```

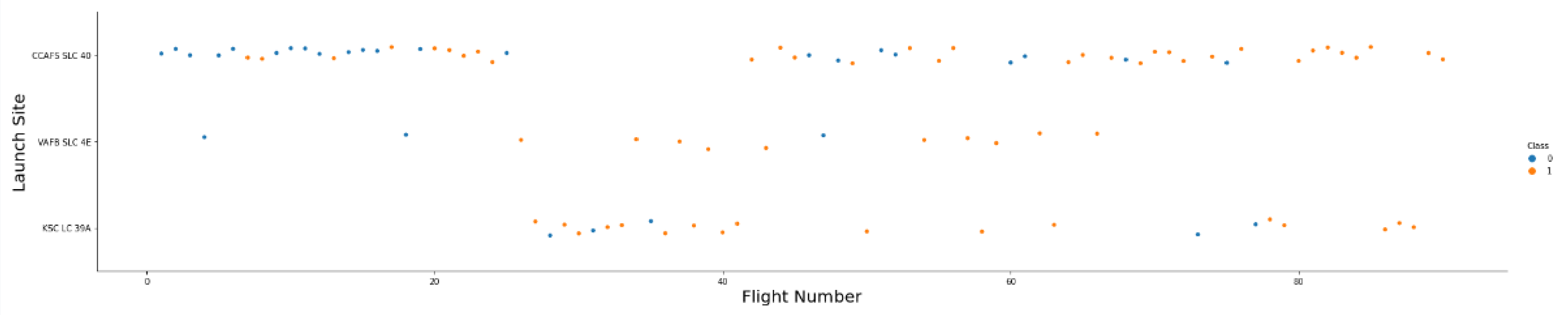
```
LogReg Score:  0.8333333333333334
SVM Score:  0.8333333333333334
Tree Score:  0.8888888888888888
KNN Score:  0.8333333333333334
```

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

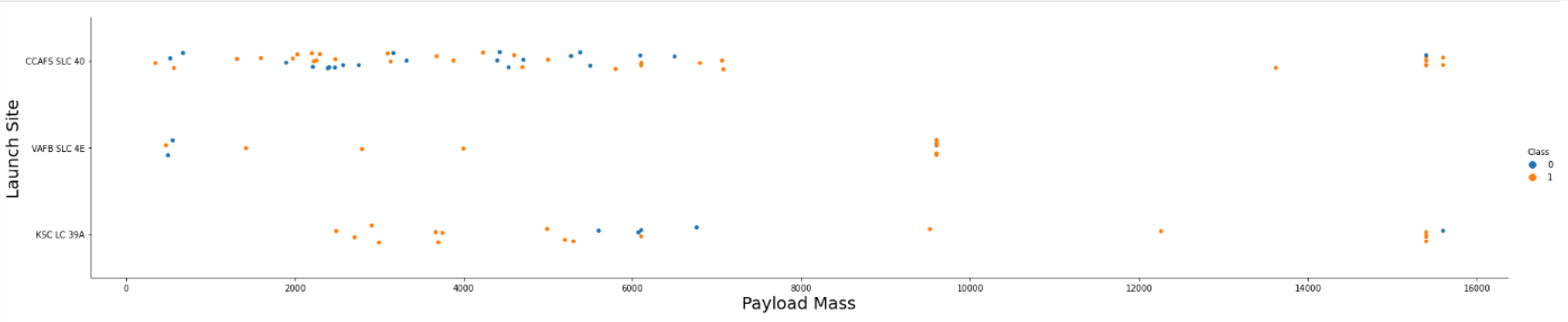
- Scatter plot of Flight Number vs. Launch Site



- Launch Sites on the Y axis and Flight Number on the X axis, Blue dots are failed landings and Orange dots are successful landings.
- You can see that KSC LC-39A has the best record of success
- Additionally, you can see CCAFS SLC 40 has improved its success over time

Payload vs. Launch Site

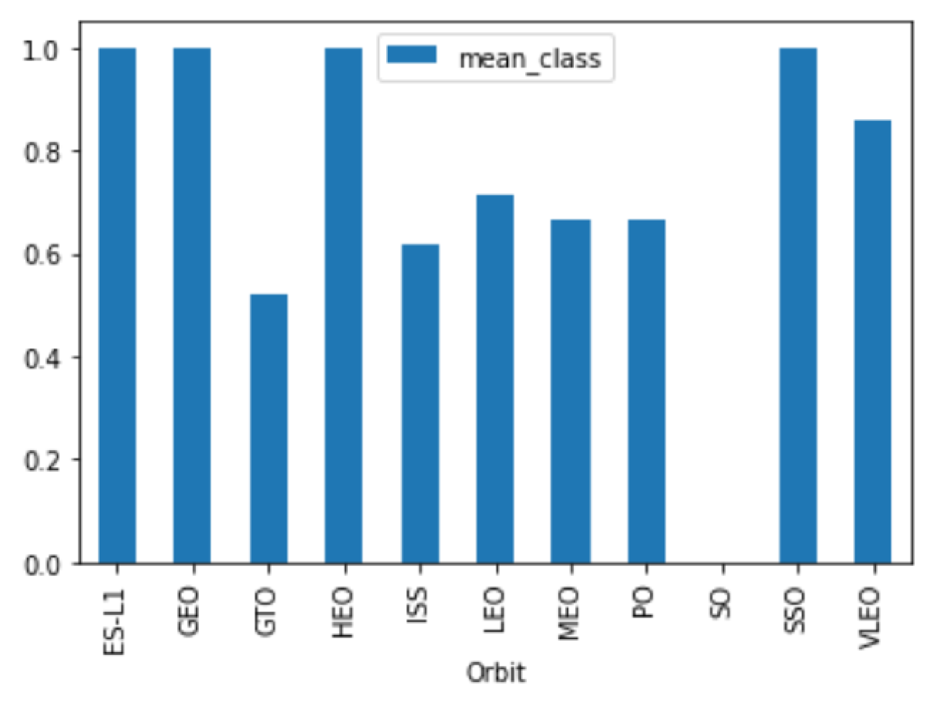
- Scatter plot of Payload vs. Launch Site



- CCAF5 SLC 40 has a better track record with heavier payload
- VAFB SLC 4E's failures happed with very light payloads
- Most success has happened with payloads from 8000 to16000 kg

Success Rate vs. Orbit Type

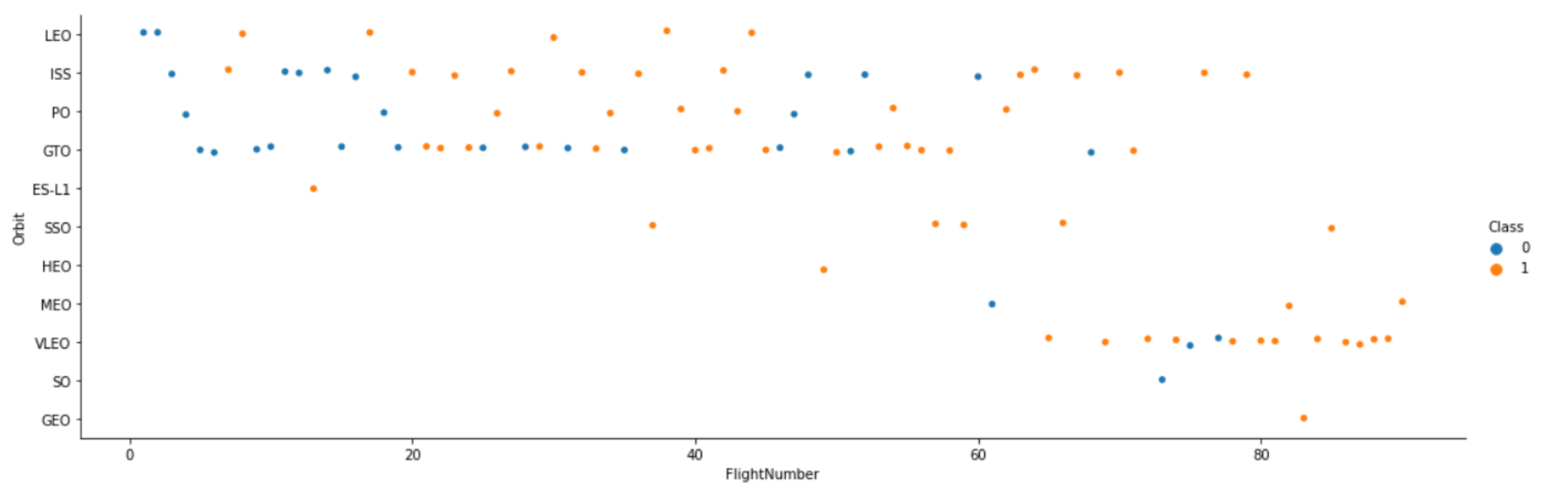
- Bar chart for the success rate of each orbit type



- ESL1, GEO, HEO, and SSO have 100% success
- VLEO is close second with 90% success
- LEO is in third with 70% success
- GTO performed the worse with 50% success
- Imagine flipping a coin to see if your multi million-dollar rocket would land!

Flight Number vs. Orbit Type

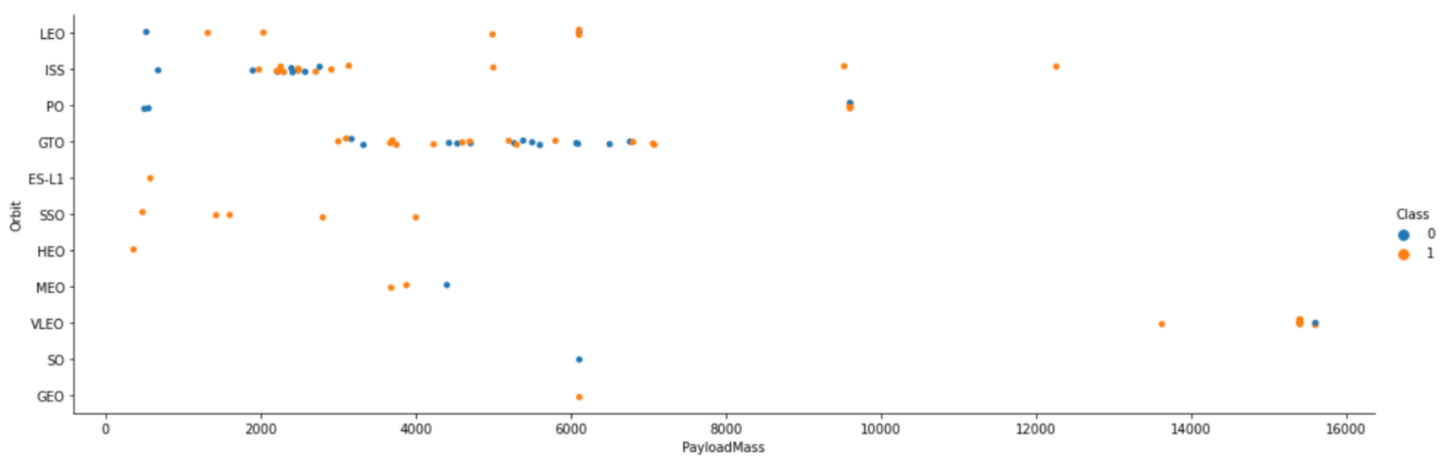
- Scatter point of Flight number vs. Orbit type



- First few launches were in LEO and ISS orbits and failed
- Launches in LEO and ISS orbits had a better success rate with time
- VLEO, SO, GEO, where only attempted later in the program

Payload vs. Orbit Type

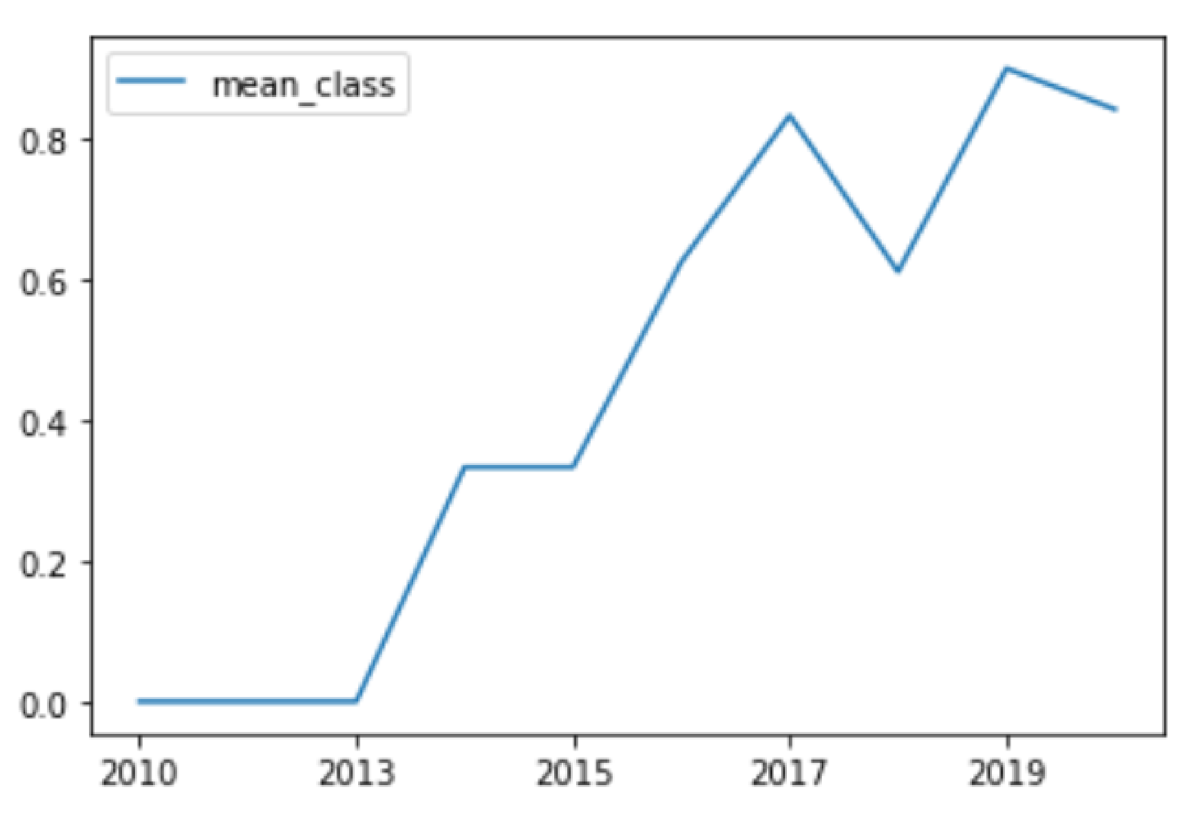
- Scatter point of payload vs. orbit type



- GTO orbit landing success does not seem to be tied to payload mass
- ES-L1, SSO, HEO, and MEO launches took place with less than 7000kg of payload
- VLEO launches saw the highest payload masses of the program

Launch Success Yearly Trend

- Line chart of yearly average success rate
- The SpaceX program started in 2010
- In 2015 the success rate was less than 40%
- 2017 the success rate jumped to over 80%
- There was a slight dip in success in 2018
- 2019 rebounded with a nearly perfect landing success rate



All Launch Site Names

- In the data set there are four unique launch sites
- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC-4E
- select distinct(LAUNCH_SITE) FROM SPACEXTBL
- The above query returns the distinct launch sites from the data set.

Launch Site Names Begin with 'KSC'

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-03-16	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

- SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'KSC%' limit 5
- Here are 5 records of launches from launch sites that start with KSC

Total Payload Mass

- select sum(payload_mass__kg_) from spacextbl where customer = 'NASA (CRS)'
- The above query sums the total payload mass of all rockets where the customer was NASA
- The total payload is 45,596 kg

Average Payload Mass by F9 v1.1

- select avg(payload_mass__kg_) from spacextbl where booster_version = 'F9 v1.1’
- The above query calculates the average payload carried by F9 v1.1 boosters
- The average payload is 2,928 kg for the F9 v1.1 boosters

First Successful Ground Landing Date

- select min(date) from spacextbl where landing__outcome = 'Success (drone ship)'
- The above query finds the first date that SpaceX performed a successful drone ship landing
- That date was 04-08-2016

Successful Drone Ship Landing with Payload between 4000 and 6000

```
select booster_version from spacextbl
where
    landing__outcome = 'Success (ground pad)'
and payload_mass__kg_ > 4000
and payload_mass__kg_ < 6000
```

booster_version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

- The above images show the query to get the booster version that have landed on a drone ship that launched with more than 4,000kg of payload and less than 6,000kg of payload.

Total Number of Successful and Failure Mission Outcomes

- select mission_outcome, count(*) as count from spacextbl group by mission_outcome
- The above query retrieves the counts of the different mission outcomes (this is not the same as landing outcomes of the booster)

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

Boosters Carried Maximum Payload

- select booster_version from spacextbl where
  payload_mass__kg_ =
  (select max(payload_mass__kg_) from spacextbl)
- The above query retrives the booster versions that have carried the maximum payload. It uses a sub query in the predicate to get the max payload.

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

2017 Launch Records

```sql
select
    monthname(date) as Month,
    landing__outcome,
    booster_version,
    launch_site
from spacextbl
where
    landing__outcome = 'Success (ground pad)'
    and year(date) = 2017
```

MONTH	landing__outcome	booster_version	launch_site
February	Success (ground pad)	F9 FT B1031.1	KSC LC-39A
May	Success (ground pad)	F9 FT B1032.1	KSC LC-39A
June	Success (ground pad)	F9 FT B1035.1	KSC LC-39A
August	Success (ground pad)	F9 B4 B1039.1	KSC LC-39A
September	Success (ground pad)	F9 B4 B1040.1	KSC LC-39A
December	Success (ground pad)	F9 FT B1035.2	CCAFS SLC-40

- The above images show the query and successful landings on ground
- The columns where specifically selected to give a summary

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

```
select landing__outcome, count(*) as cnt from spacextbl
where landing__outcome like 'Success%'
group by landing__outcome
order by 2
```

landing__outcome	cnt
Success (ground pad)	9
Success (drone ship)	14
Success	38

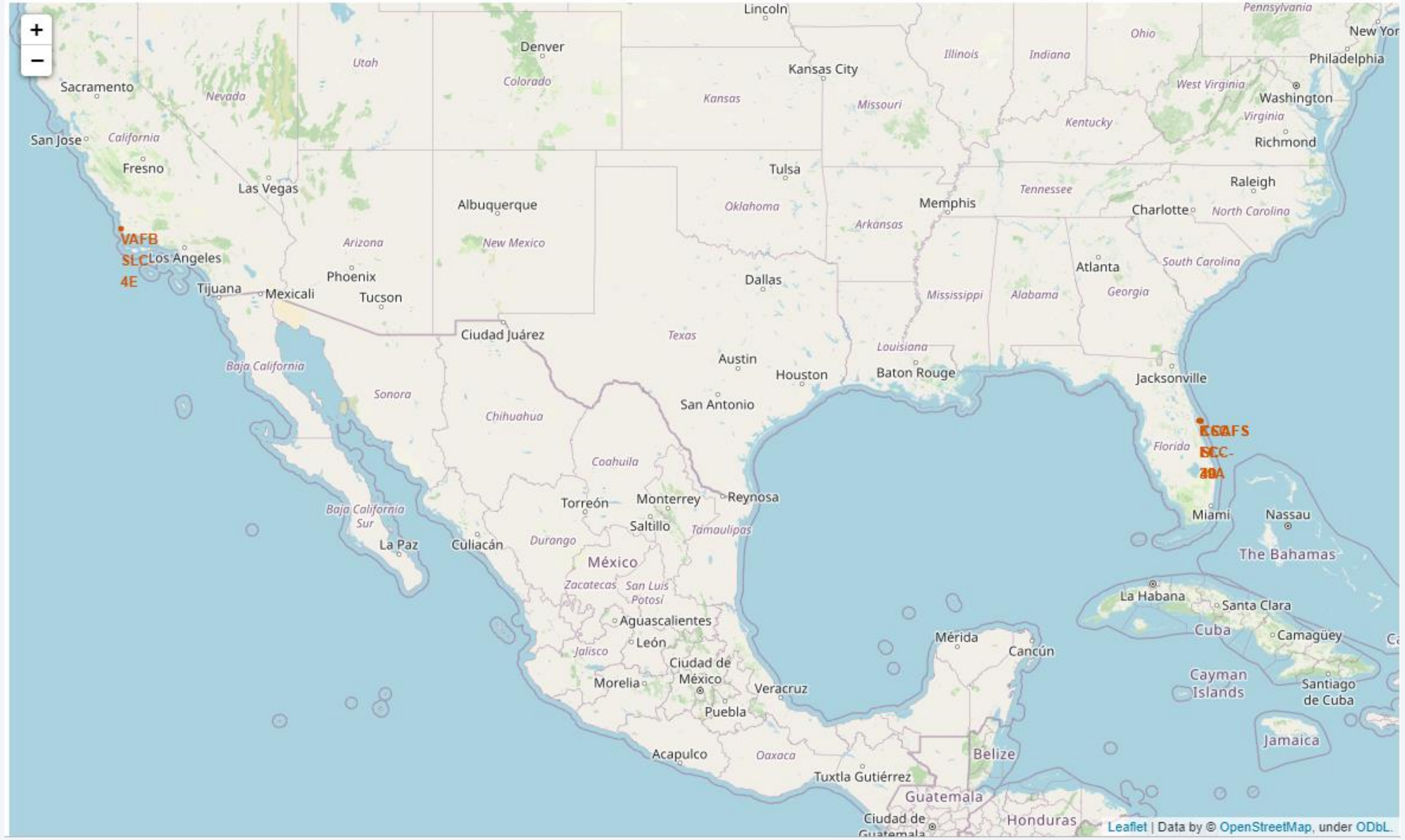
- The above images show the query and the successful landings ranked by landing outcome
- Group by must be used when an aggregate function is used such as count(*)

Section 3

Launch Sites Proximities Analysis

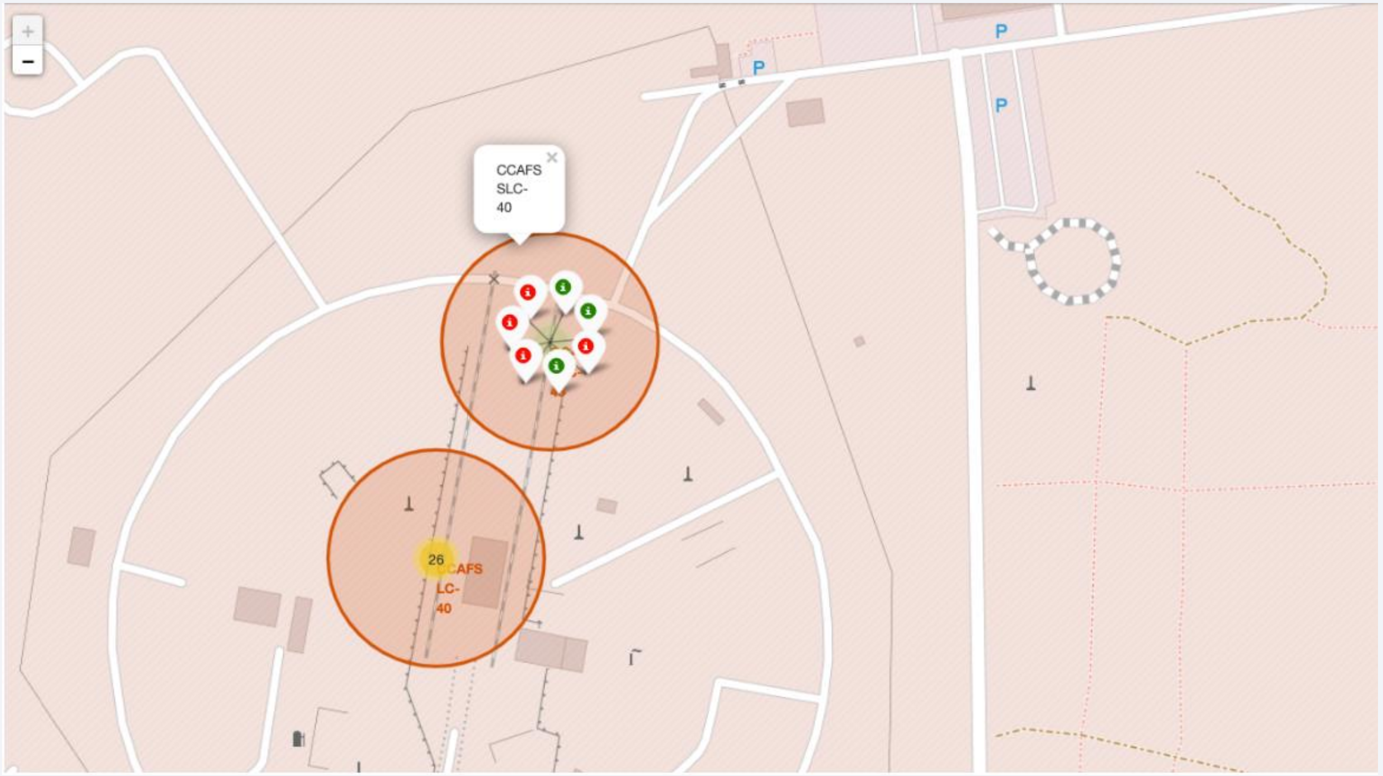
Map of Launch Sites

- I found that all SpaceX launch sites are located on either the West or East Cost of the United States of America



Colored Labels Based On Landing Success

- Clustering multiple records and coloring them based on landing success can give more details on a map then if we placed all icons on the exact position of the launch



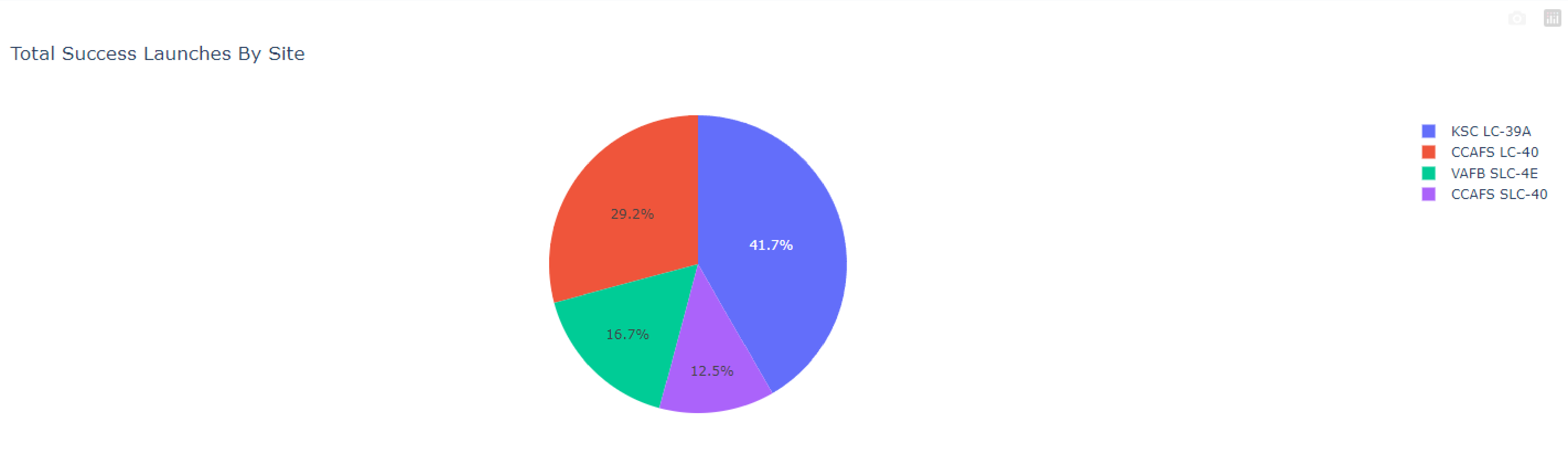
Launch Sites and Proximities

- I found it interesting that launch sites seem to be in extremely close proximity to railroads. Most likely to transport rockets and supplies.
- I also noticed that launch sites seem to avoid crossing major highways on their short path over the ocean, in some cases less then 1 km away.

Section 4

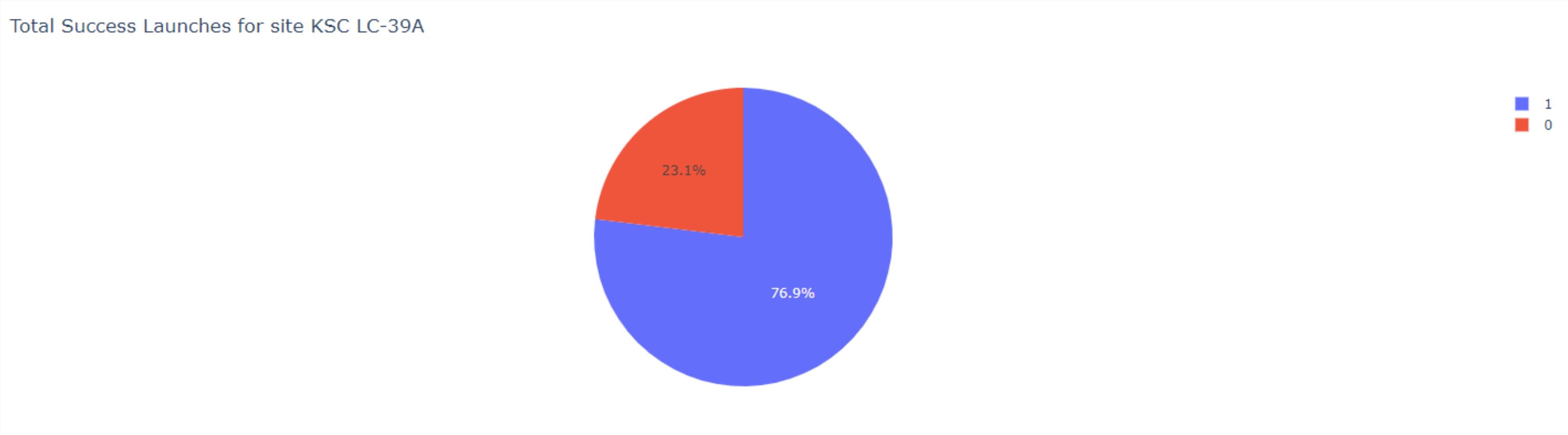
Build a Dashboard with Plotly Dash

Successful Landings for all Sites



- KSC LC-39A was the most successful of all the sites
- CCAFS LC-40 was second
- CCAFS SLC-40 was the least successful

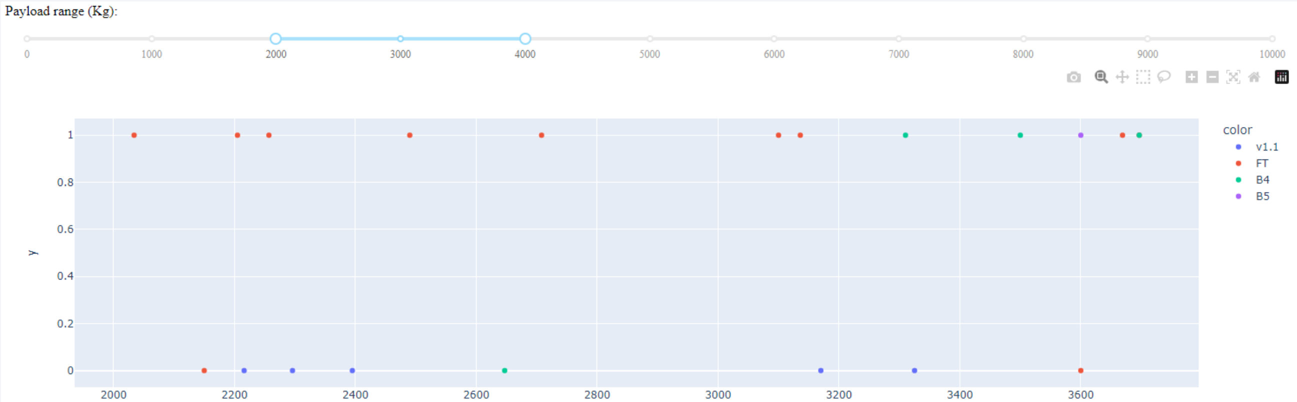
Most Successful Launch Site



- KSC LC-39A was the most successful launch site
- About 77% of launches ended with a successful landing
- Only about 23% of launched boosters failed to land

Launch Success vs Payload

- Launch success for all sites and payloads
- A mix of success and failure can be seen

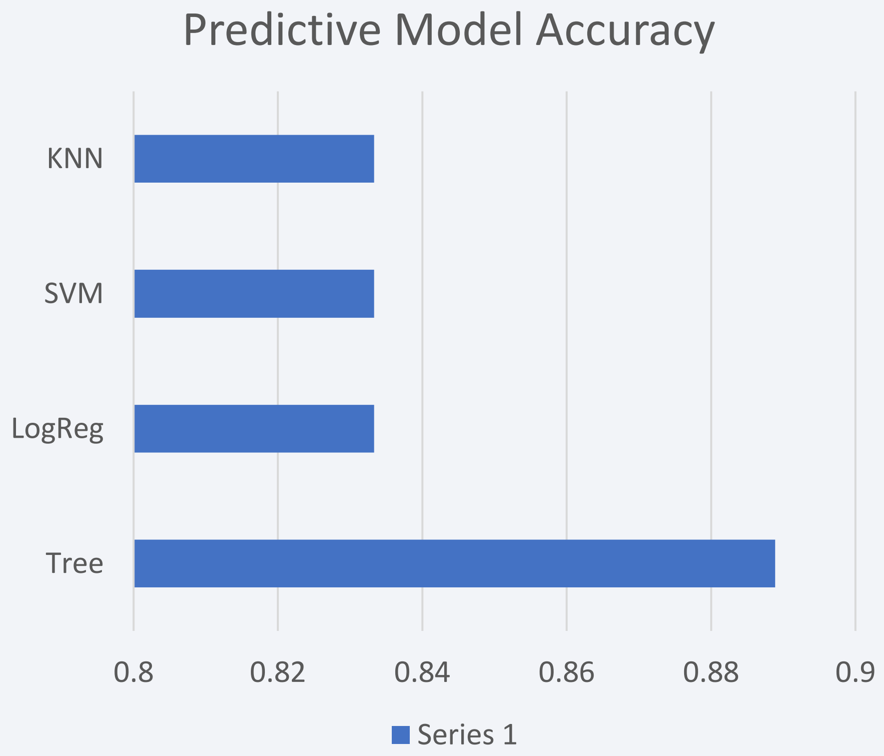


- Launch success for all sites and payload between 2,000 and 4,000 kg
- This was the most successful range of payloads

Section 5

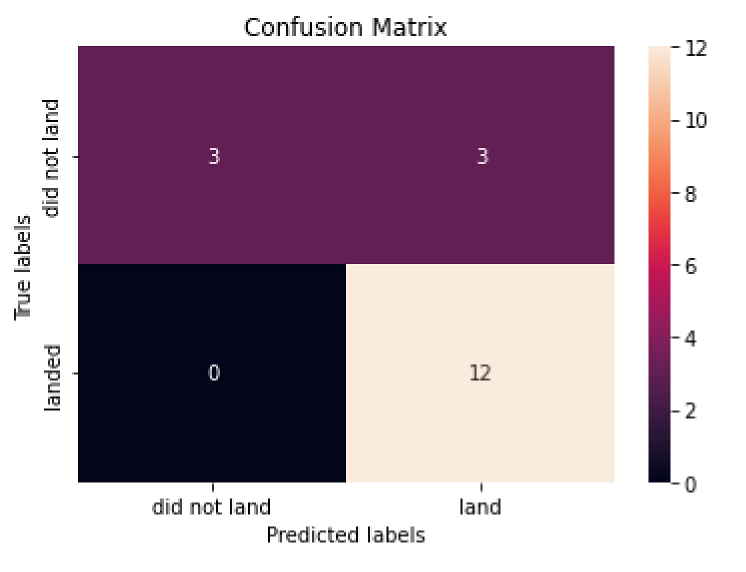
Predictive Analysis (Classification)

Classification Accuracy



- The Tree Model preformed the best when the test data was applied.
- Greater than 88% accuracy could be achieved using the tree model

Tree Confusion Matrix



- While the tree model performed the best in my case it isn't perfect.
- The tree was good at predicting successful landings getting 12 correct and 3 incorrect.
- The tree was bad at predicting failed landings, it miss all three failures.

Conclusions

- Successful landings can be predicted with a limited number of false positives and false negatives
- Launch sites, payload, flight number, and orbit are good predicting factors
- Predictions could improve with a larger sample of known outcomes
- Models vary in predictive accuracy
- Predictions could improve with data sets specific to the one of the predicting factors. Such as models made per Launch Site, Payload, Orbit etc.

Thank you!