IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Morten Ehari
08.10.2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data scraping from API
  - Data scraping from web
  - Data wrangling
  - Exploratory data analysis with SQL
  - Exploratory data analysis with data visualisation
  - Interactive visual analytics with Folium
  - Machine learning prediction
- Summary of all results
  - Exploratory data analysis result
  - Predictive analytics result

Introduction

- Project background and context
  - SpaceX advertises Falcon 9 rocket launches on it website with a cost of 62 million dollars
  - If we can determine if the first stage will land, we can determine the cost of a launch
  - This information is useful if we want to bid against SpaceX for a rocket launch

- Problems you want to find answers
  - What factors determine if the rocket will land successfully?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What are the operating conditions for a successful landing program?

Section 1

Methodology

Methodology

Executive Summary

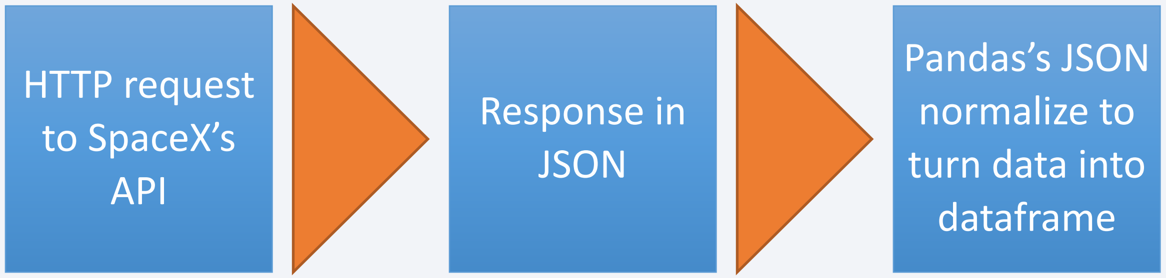
- Data collection methodology:
    - Data was collected through:
        - SpaceX API
        - Web scraping Wikipedia
- Perform data wrangling
    - One-hot encoding was applied to categorical features
- Perform exploratory data analysis (EDA) using visualization and SQL
    - Discovering new patters in the data with visualisation
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Classification machine learning models were created and evaluated (using confusion matrices and so on…)

Data Collection

- Data came from two origins - SpaceX’s API and SpaceX’s wikipedia page
- SpaceX’s API information includes data about launches - launch and landing specifics, payload delivery result, landing outcome and so on…
- Data collected from SpaceX’s wikipedia page include payload mass, orbit, booster version, customer and so on….

Data Collection – SpaceX API

- Data was collected using REST calls on SpaceX's API
- https://github.com/0x1F0FFF/DS-ML-Capstone-Project/blob/main/1.%20SpaceX%20-%20Collecting%20the%20data.ipynb



Collected data includes:
1. Booster name
2. Launchpad location
3. Payload mass
4. Core information
5. Outcome
And more…

Data Collection - Scraping

- Process work flow can be seen in the flowchart

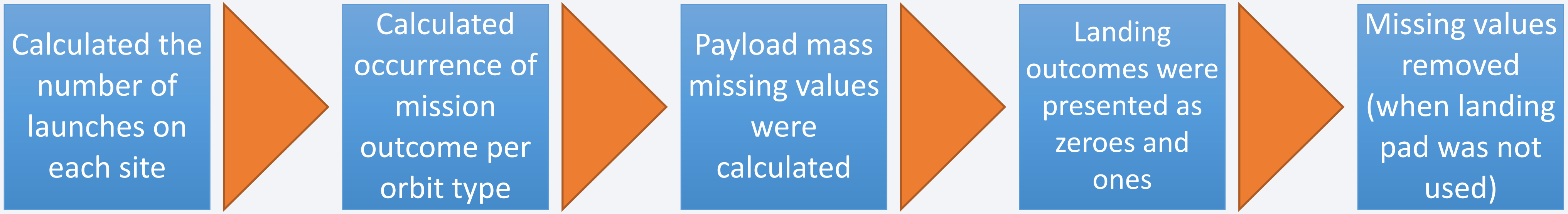
- https://github.com/0x1F0FFF/DS-ML-Capstone-Project/blob/main/1.%20SpaceX%20-%20Web%20scraping.ipynb

Information extracted:

1. Data and time
2. Booster version
3. Launch site
4. Payload
5. Payload mass
6. Orbit
7. Launch outcome
8. Booster landing outcome
And more…

Data Wrangling

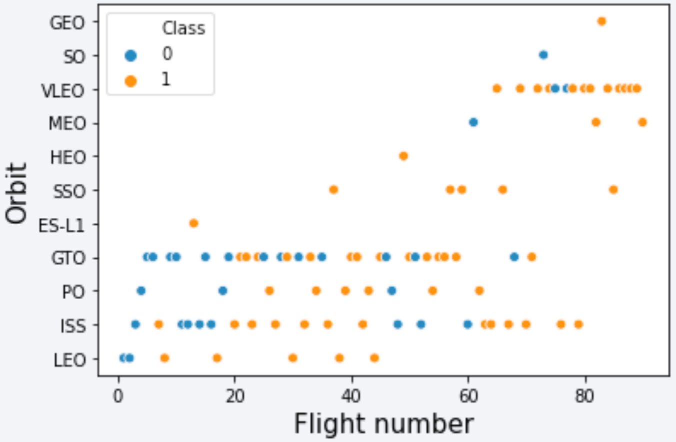
- Data analysis was done to find patterns in the data
- Wrangling overall scheme looked like this:



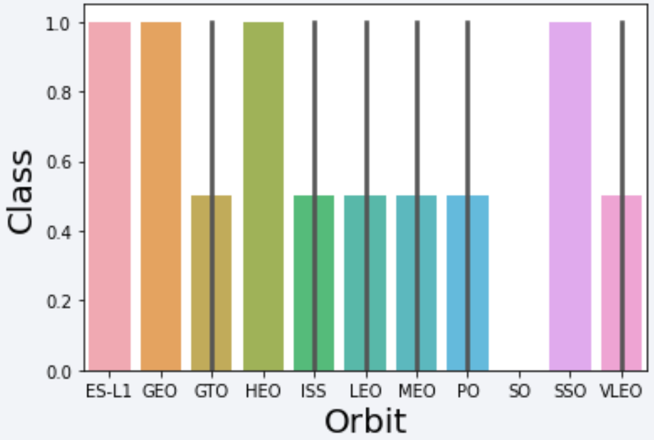
- https://github.com/0x1F0FFF/DS-ML-Capstone-Project/blob/main/2.%20SpaceX%20-%20Data%20wrangling.ipynb

EDA with Data Visualization

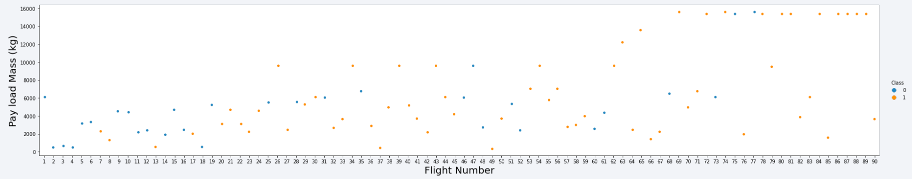
- Charts were chosen by thinking what charts would best display the trend
- Some of them are shown here:



Relationship between flight number and orbit type



Relationship between success rate of each orbit type



Relationship between flight number and payload mass

- https://github.com/0x1F0FFF/DS-ML-Capstone-Project/blob/main/4.%20SpaceX%20-%20Exploring%20and%20preparing%20data.ipynb

EDA with SQL

- SQL queries performed:
  - Displayed names of the unique launch sites in the space mission
  - Displayed five records, where launch sites begin the string “CCA”
  - Displayed the total payload mass carried by boosters launched by NASA (CRS)
  - Displayed the average payload mass carried by booster version F9 v1.1
  - Displayed the date when the first successful landing outcome in ground pad was achieved
  - Displayed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Displayed the total number of successful and failure mission outcomes
  - Displayed the names of the booster versions which have carried the maximum payload mass
  - Displayed the failed landing outcomes in drone ship, their booster version and launch site names in year 2015
  - Ranked the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order
- https://github.com/0x1F0FFF/DS-ML-Capstone-Project/blob/main/5.%20SpaceX%20-%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Added map objects include:
    - Marked launch sites
    - Success and failed launches for each site
    - Distances between a launch site to its proximities
- Used functions include Marker(), Circle(), Icon() and PolyLine()
- Objects were added on the map to analyse trends based on launch site locations
- https://github.com/0x1F0FFF/DS-ML-Capstone-Project/blob/main/5.%20SpaceX%20-%20Launch%20sites%20locations%20analysis%20with%20Folium.ipynb

Predictive Analysis (Classification)

Build the model

Standardised the data
Split into train and test sets
Build model

Evaluate the model

Calculate model accuracy

Calculate the confusion matrices

Use the optimal model

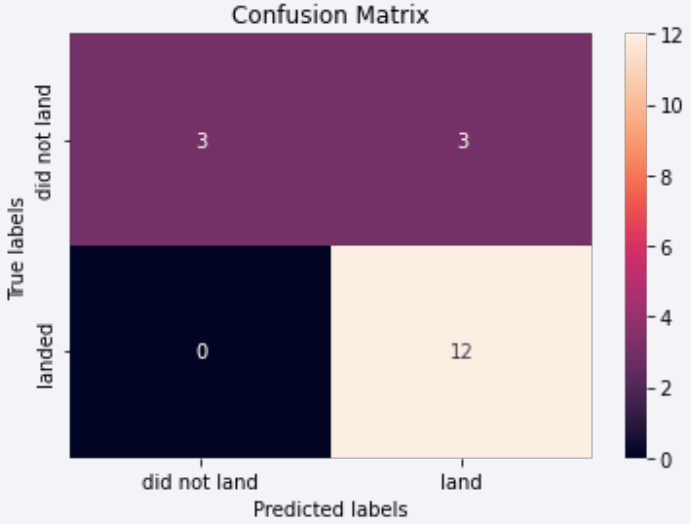
Find the model with the highest accuracy

Find hyper parameters

- https://github.com/0x1F0FFF/DS-ML-Capstone-Project/blob/main/5.%20SpaceX%20-%20Machine%20learning%20prediction.ipynb

Results

- Exploratory data analysis results:
  - Heavier payloads had worse results than low weight ones
  - Over time launch success rates have gotten better due to technological advancement
  - Logistic regression, Support vector machine and K-nearest neighbors models were most accurate in prediction

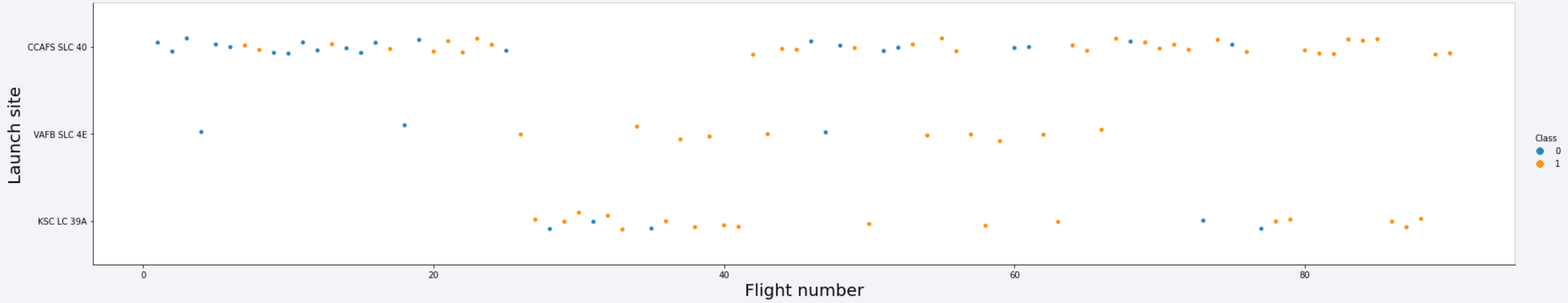


KNN confusion matrix

Section 2

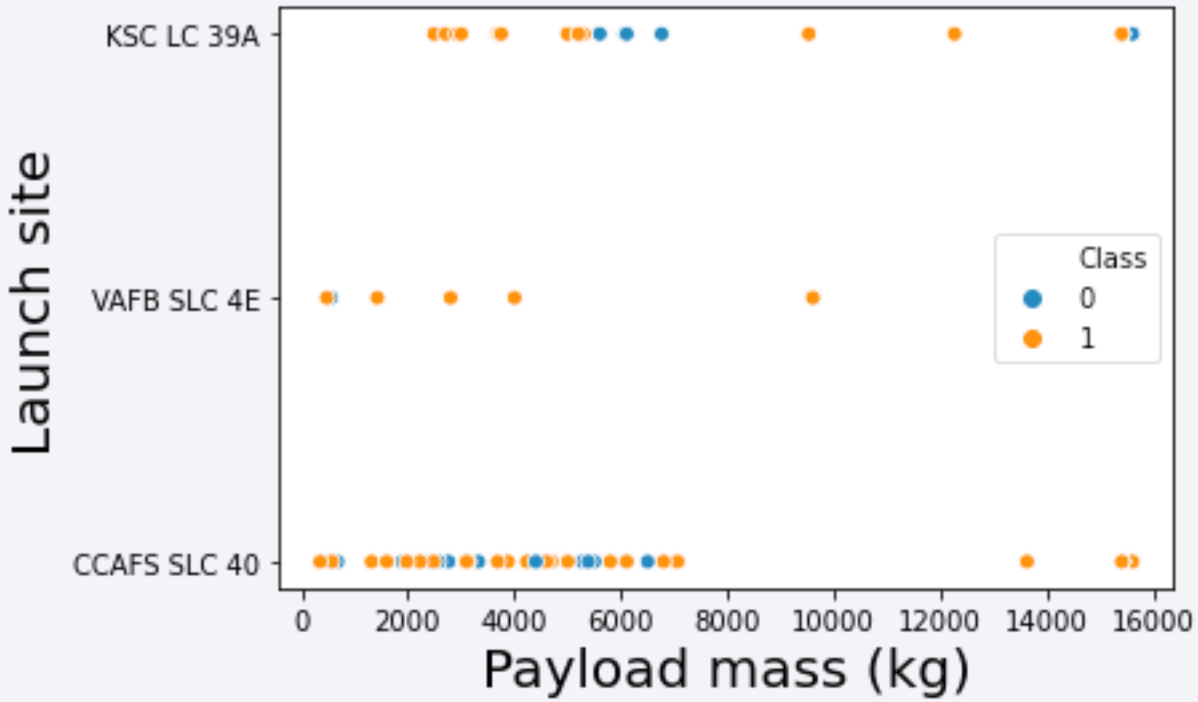
Insights drawn from EDA

Flight Number vs. Launch Site



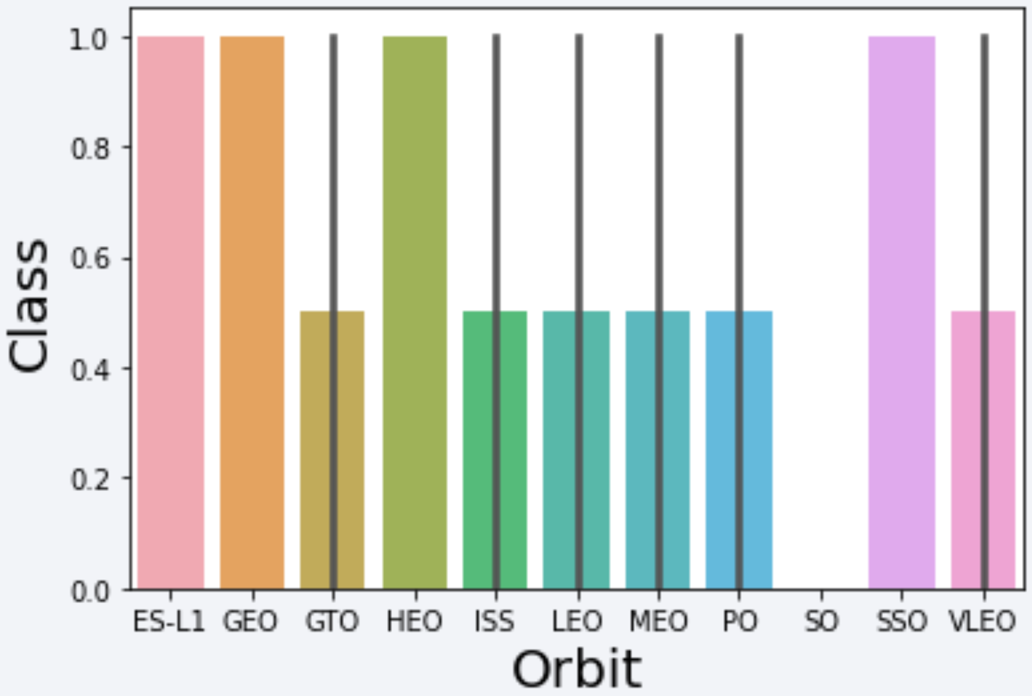
- As displayed, no matter the launch site, with increase of flight number, the success rate has increased and CCAFS SLC 40 success rate is higher.

Payload vs. Launch Site



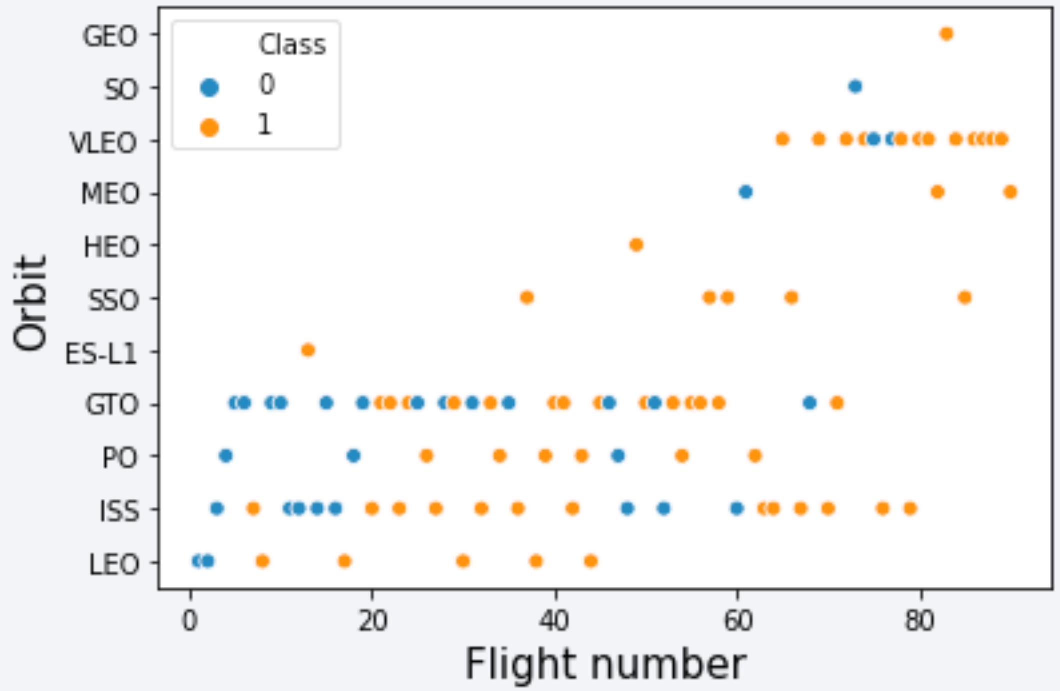
- Lighter payloads have been launched at CCAFS SLC 40 - which could explain previous slide's chart.

Success Rate vs. Orbit Type



- ES-L1, GEO, HEO and SSO have the highest success rate

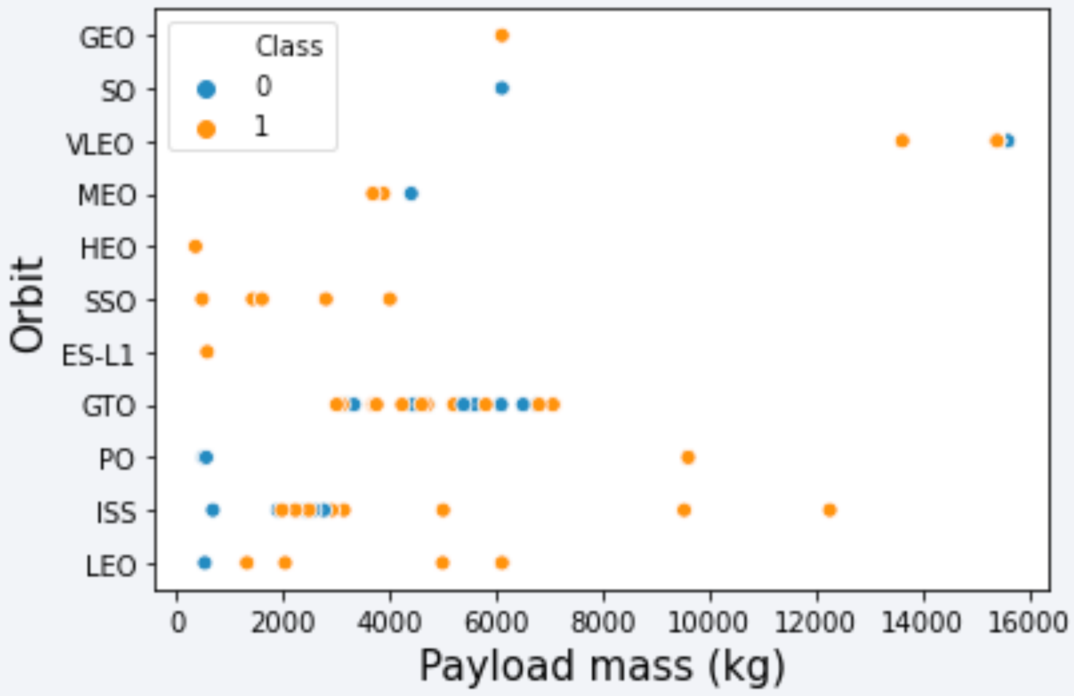
Flight Number vs. Orbit Type



- Increase in VLEO launches can be observed

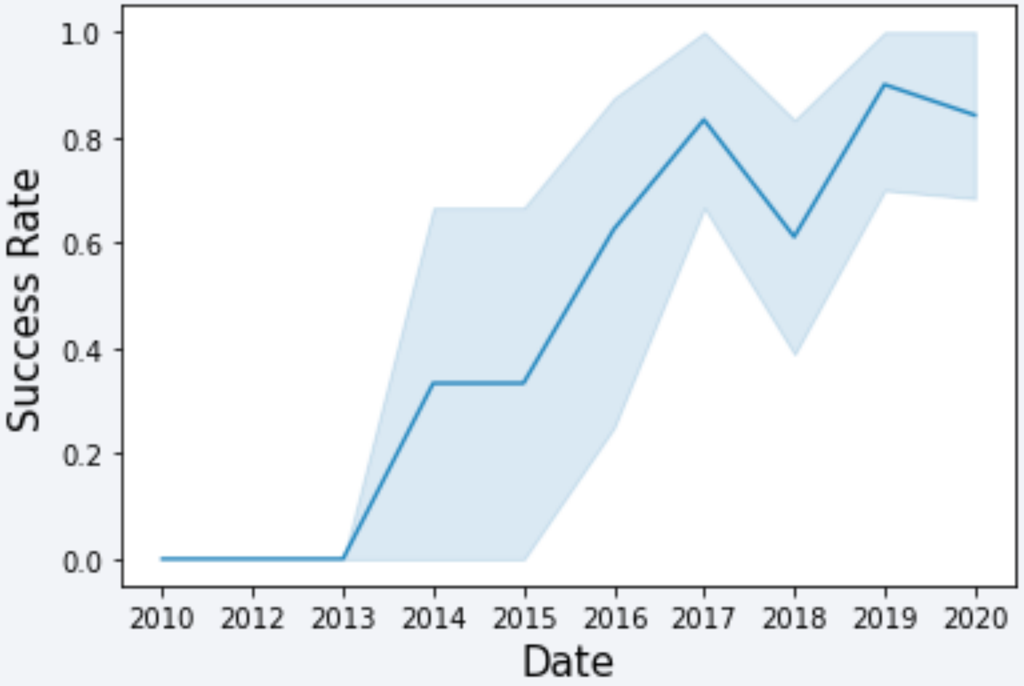
Payload vs. Orbit Type

- Two correlations can be observed here:
  - Payload around 2000 and ISS
  - Payload around 4000 - 8000 and GTO



Launch Success Yearly Trend

- Trend seen between years 2013 and 2017, where success rate has climbed a lot. Overall success rate has increased a lot.



All Launch Site Names

- Unique values were found using DISTINCT

```
sqlite> SELECT DISTINCT Launch_Site FROM SPACEX;
"CCAFS LC-40"
"VAFB SLC-4E"
"KSC LC-39A"
"CCAFS SLC-40"
sqlite>
```

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

Launch Site Names Begin with 'KSC'

- Find 5 records where launch sites' names start with `KSC`
- SELECT * FROM SPACEX WHERE Launch_Site LIKE '%KSC%' LIMIT 5;

```
sqlite> SELECT * FROM SPACEX WHERE Launch_Site LIKE "%KSC%" LIMIT 5;
19-02-2017,14:39:00,"F9 FT B1031.1","KSC LC-39A","SpaceX CRS-10",2490,"LEO (ISS)","NASA (CRS)",Success,"Success (ground pad)"
16-03-2017,06:00:00,"F9 FT B1030","KSC LC-39A","EchoStar 23",5600,GTO,EchoStar,Success,"No attempt"
30-03-2017,22:27:00,"F9 FT  B1021.2","KSC LC-39A",SES-10,5300,GTO,SES,Success,"Success (drone ship)"
01-05-2017,11:15:00,"F9 FT B1032.1","KSC LC-39A",NROL-76,5300,LEO,NRO,Success,"Success (ground pad)"
15-05-2017,23:21:00,"F9 FT B1034","KSC LC-39A","Inmarsat-5 F4",6070,GTO,Inmarsat,Success,"No attempt"
sqlite>
```

Total Payload Mass

- To find out the total payload mass carried by NASA we can use this query:
- SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEX WHERE CUSTOMER = “NASA (CRS)”;
- The keyword SUM adds it all up

Total payload mass by NASA (CRS)
45596

```
sqlite> SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEX WHERE CUSTOMER = "NASA (CRS)";
45596
sqlite>
```

Average Payload Mass by F9 v1.1

- To get the average payload mass carried by booster version F9 v1.1 this query can be used:
  - SELECT AVG(Payload_Mass__KG_) AS Avg_PayloadMass FROM SPACEX WHERE Booster_Version = “F9 v1.1”;

  - The average payload in this context is 2928,4 kilograms.

```
sqlite> SELECT AVG(PAYLOAD_MASS__KG_) AS Avg_PayloadMass FROM SPACEX WHERE Booster_Version = "F9 v1.1";
2928.4
sqlite>
```

First Successful Drone Ship Landing Date

- To find the date of the first successful landing outcome on drone ship we can use this query:
    - SELECT MIN(DATE) AS “First Successful Landing Outcome on Drone Ship” FROM SPACEX WHERE “Landing _Outcome” = “Success (drone ship)”;

- Which returns us the date 06-05-2016

```
sqlite> sqlite> SELECT MIN(DATE) AS "First Successful Landing Outcome on Drone Ship" FROM SPACEX WHERE "Landing _Outcome" = "Success (drone ship)";
06-05-2016
sqlite> 
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- To get the list of names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 we can use this query:
  - SELECT Booster_Version FROM SPACEX WHERE “Landing _Outcome” = “Success (drone ship) AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
- The result is:

```
sqlite> SELECT Booster_Version FROM SPACEX WHERE "Landing _Outcome" = "Success (drone ship)" AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
"F9 FT B1022"
"F9 FT B1026"
"F9 FT  B1021.2"
"F9 FT B1038.1"
"F9 FT  B1031.2"
sqlite>
```

Total Number of Successful and Failure Mission Outcomes

- To calculate the total number of successful and failed mission outcomes we can use these two queries:
  - SELECT COUNT(Mission_Outcome) AS “Successful Mission” FROM SPACEX WHERE Mission_Outcome LIKE “Success%”;
  - SELECT COUNT(Mission_Outcome) AS “Failure Mission” FROM SPACEX WHERE Mission_Outcome LIKE “Failure%”;

```
sqlite> SELECT COUNT(Mission_Outcome) AS "Successful Mission" FROM SPACEX WHERE Mission_Outcome LIKE "Success%";
100
sqlite> SELECT COUNT(Mission_Outcome) AS "Failure Mission" FROM SPACEX WHERE Mission_Outcome LIKE "Failure%";
1
sqlite>
```

- Essentially these two queries are the same.

Boosters Carried Maximum Payload

- To list the booster that have carried the maximum payload mass we can use this query:

```
sqlite> SELECT DISTINCT BOOSTER_VERSION AS "Booster Versions which carried the Maximum Payload Mass" FROM SPACEX WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX);
"F9 FT B1029.1"
"F9 FT B1036.1"
"F9 B4 B1041.1"
"F9 FT  B1036.2"
"F9 B4  B1041.2"
"F9 B5B1048.1"
"F9 B5 B1049.2"
sqlite>
```

- As you can see, at this moment, seven boosters have carried the maximum payload mass.

2015 Launch Records

- List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

```
sqlite> SELECT * FROM SPACEX WHERE "Landing _Outcome" like "Success%" AND (DATETIME between "2017-01-01" and "2017-12-31") order by date desc;
43,30-10-2017,2017-10-30,19:34:00,"F9 B4 B1042.1","KSC LC-39A","Koreasat 5A",3500,GTO,"KT Corporation",Success,"Success (drone ship)"
31,30-03-2017,2017-03-30,22:27:00,"F9 FT  B1021.2","KSC LC-39A",SES-10,5300,GTO,SES,Success,"Success (drone ship)"
36,25-06-2017,2017-06-25,20:25:00,"F9 FT B1036.1","VAFB SLC-4E","Iridium NEXT 2",9600,LEO,"Iridium Communications",Success,"Success (drone ship)"
39,24-08-2017,2017-08-24,18:51:00,"F9 FT B1038.1","VAFB SLC-4E",Formosat-5,475,SSO,NSPO,Success,"Success (drone ship)"
35,23-06-2017,2017-06-23,19:10:00,"F9 FT  B1029.2","KSC LC-39A",BulgariaSat-1,3669,GTO,Bulsatcom,Success,"Success (drone ship)"
29,19-02-2017,2017-02-19,14:39:00,"F9 FT B1031.1","KSC LC-39A","SpaceX CRS-10",2490,"LEO (ISS)","NASA (CRS)",Success,"Success (ground pad)"
44,15-12-2017,2017-12-15,15:36:00,"F9 FT  B1035.2","CCAFS SLC-40","SpaceX CRS-13",2205,"LEO (ISS)","NASA (CRS)",Success,"Success (ground pad)"
38,14-08-2017,2017-08-14,16:31:00,"F9 B4 B1039.1","KSC LC-39A","SpaceX CRS-12",3310,"LEO (ISS)","NASA (CRS)",Success,"Success (ground pad)"
28,14-01-2017,2017-01-14,17:54:00,"F9 FT B1029.1","VAFB SLC-4E","Iridium NEXT 1",9600,"Polar LEO","Iridium Communications",Success,"Success (drone ship)"
42,11-10-2017,2017-10-11,22:53:00,"F9 FT  B1031.2","KSC LC-39A","SES-11 / EchoStar 105",5200,GTO,"SES EchoStar",Success,"Success (drone ship)"
41,09-10-2017,2017-10-09,12:37:00,"F9 B4 B1041.1","VAFB SLC-4E","Iridium NEXT 3",9600,"Polar LEO","Iridium Communications",Success,"Success (drone ship)"
40,07-09-2017,2017-09-07,14:00:00,"F9 B4 B1040.1","KSC LC-39A","Boeing X-37B OTV-5",4990,LEO,"U.S. Air Force",Success,"Success (ground pad)"
34,03-06-2017,2017-06-03,21:07:00,"F9 FT B1035.1","KSC LC-39A","SpaceX CRS-11",2708,"LEO (ISS)","NASA (CRS)",Success,"Success (ground pad)"
32,01-05-2017,2017-05-01,11:15:00,"F9 FT B1032.1","KSC LC-39A",NROL-76,5300,LEO,NRO,Success,"Success (ground pad)"
sqlite>
```

- Since the title of the slide asked for 2015 launch records, here it is:

```
sqlite> SELECT * FROM SPACEX WHERE "Landing _Outcome" like "Success%" AND (DATETIME between "2015-01-01" and "2015-12-31") order by date desc;
19,22-12-2015,2015-12-22,01:29:00,"F9 FT B1019","CCAFS LC-40","OG2 Mission 2  11 Orbcomm-OG2 satellites",2034,LEO,Orbcomm,Success,"Success (ground pad)"
sqlite>
```

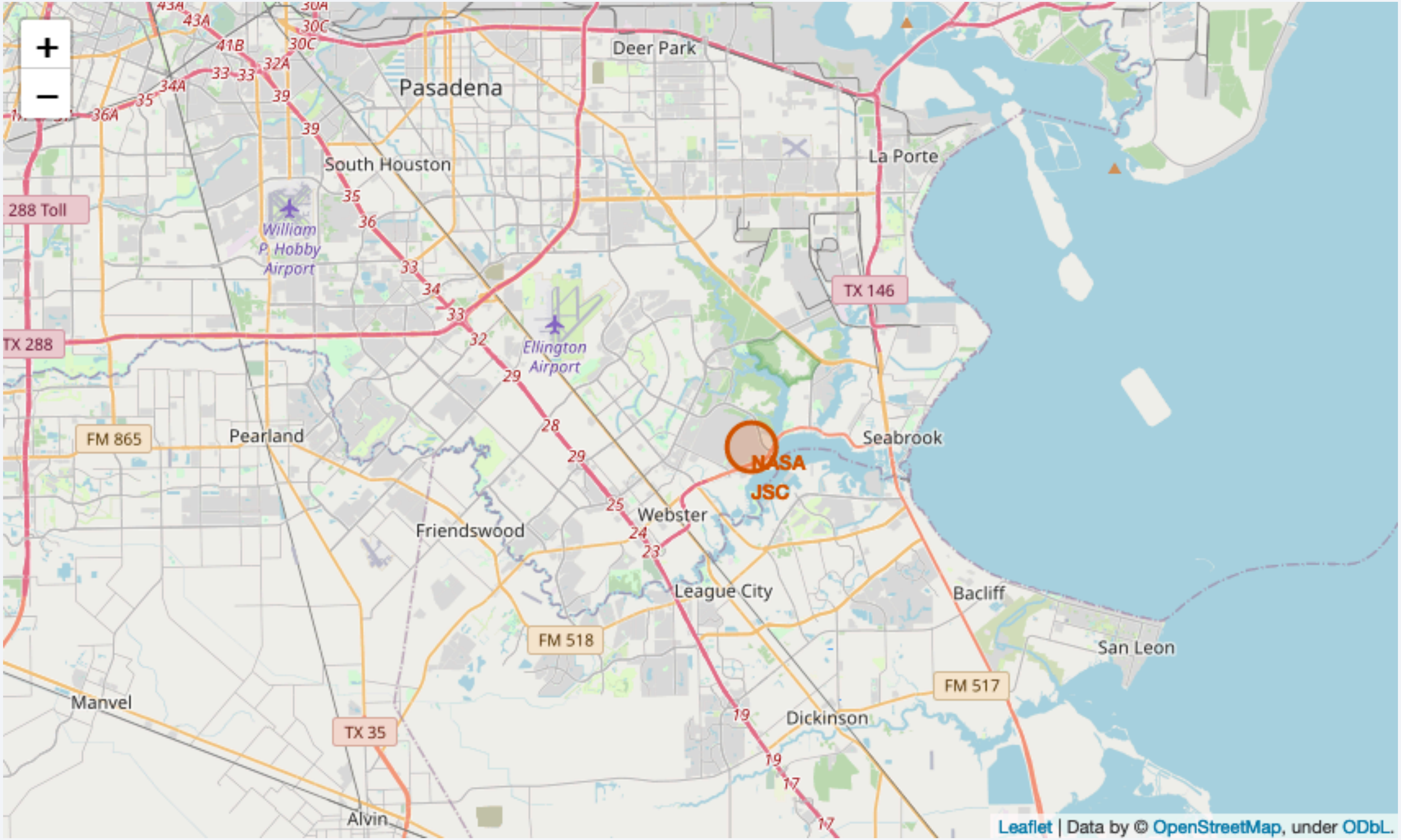
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- To rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order, we can use this query:

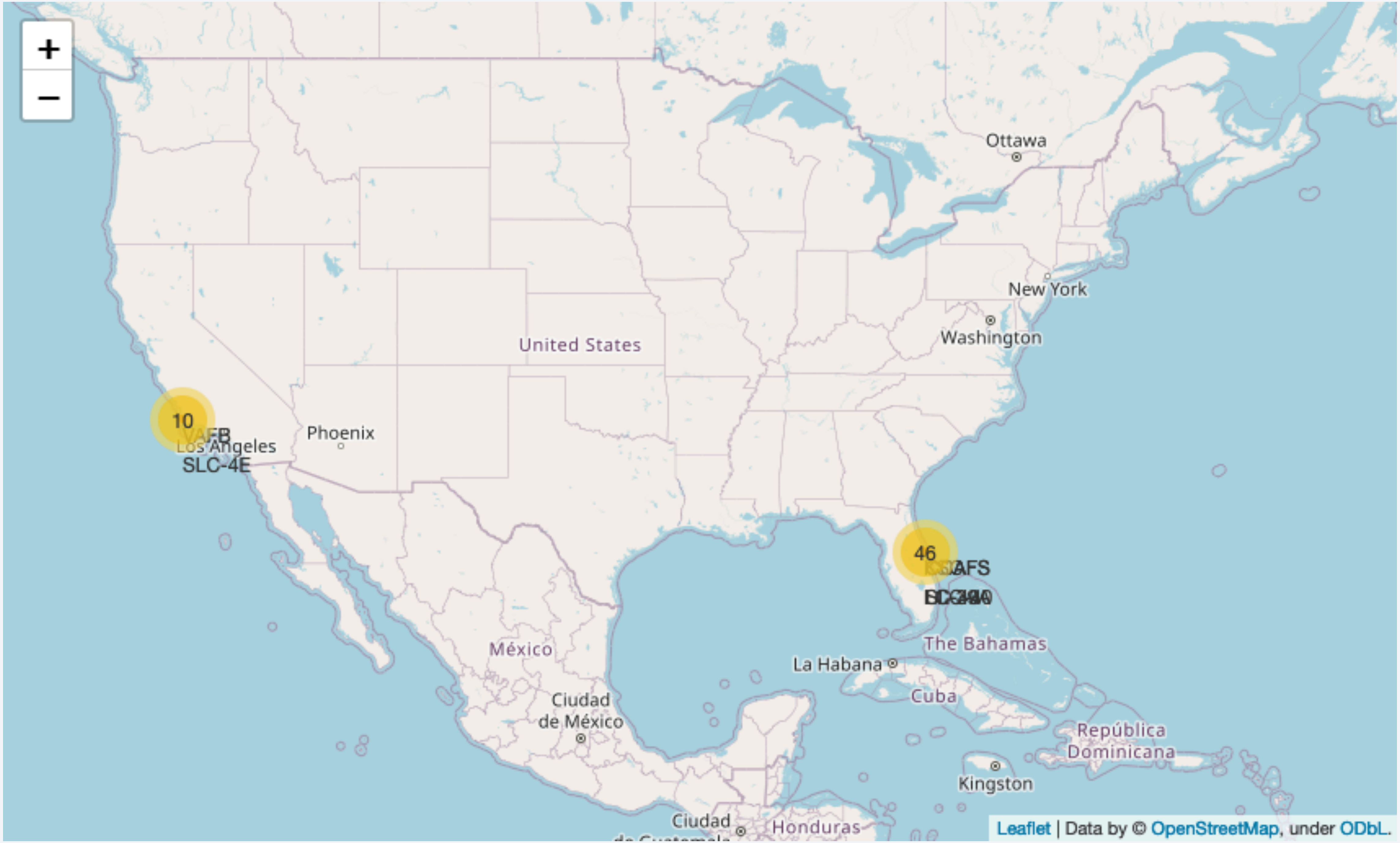
```
sqlite> SELECT * FROM SPACEX WHERE "Landing _Outcome" LIKE "Success%" and (DATETIME between "2010-06-04" AND "2017-03-20") ORDER BY DATE desc;
24,27-05-2016,2016-05-27,21:39:00,"F9 FT B1023.1","CCAFS LC-40","Thaicom 8",3100,GTO,Thaicom,Success,"Success (drone ship)"
19,22-12-2015,2015-12-22,01:29:00,"F9 FT B1019","CCAFS LC-40","OG2 Mission 2  11 Orbcomm-OG2 satellites",2034,LEO,Orbcomm,Success,"Success (ground pad)"
29,19-02-2017,2017-02-19,14:39:00,"F9 FT B1031.1","KSC LC-39A","SpaceX CRS-10",2490,"LEO (ISS)","NASA (CRS)",Success,"Success (ground pad)"
26,18-07-2016,2016-07-18,04:45:00,"F9 FT B1025.1","CCAFS LC-40","SpaceX CRS-9",2257,"LEO (ISS)","NASA (CRS)",Success,"Success (ground pad)"
27,14-08-2016,2016-08-14,05:26:00,"F9 FT B1026","CCAFS LC-40",JCSAT-16,4600,GTO,"SKY Perfect JSAT Group",Success,"Success (drone ship)"
28,14-01-2017,2017-01-14,17:54:00,"F9 FT B1029.1","VAFB SLC-4E","Iridium NEXT 1",9600,"Polar LEO","Iridium Communications",Success,"Success (drone ship)"
22,08-04-2016,2016-04-08,20:43:00,"F9 FT B1021.1","CCAFS LC-40","SpaceX CRS-8",3136,"LEO (ISS)","NASA (CRS)",Success,"Success (drone ship)"
23,06-05-2016,2016-05-06,05:21:00,"F9 FT B1022","CCAFS LC-40",JCSAT-14,4696,GTO,"SKY Perfect JSAT Group",Success,"Success (drone ship)"
sqlite>
```

Section 3

Launch Sites Proximities Analysis


Deer Park
Pasadena
South Houston
La Porte
288 Toll
William P. Hobby Airport
TX 146
Ellington Airport
TX 288
FM 865
Pearland
Seabrook
NASA JSC
Webster
Friendswood
League City
Bacliff
FM 518
San Leon
FM 517
Manvel
TX 35
Dickinson
Alvin
Leaflet | Data by © OpenStreetMap, under ODbL.

+
−
Ottawa
New York
Washington
United States
VAFB
Los Angeles
SLC-4E
Phoenix
México
Ciudad de México
The Bahamas
La Habana
Cuba
República Dominicana
Kingston
Ciudad de Guatemala
Honduras
Nicaragua
Leaflet | Data by © OpenStreetMap, under ODbL.

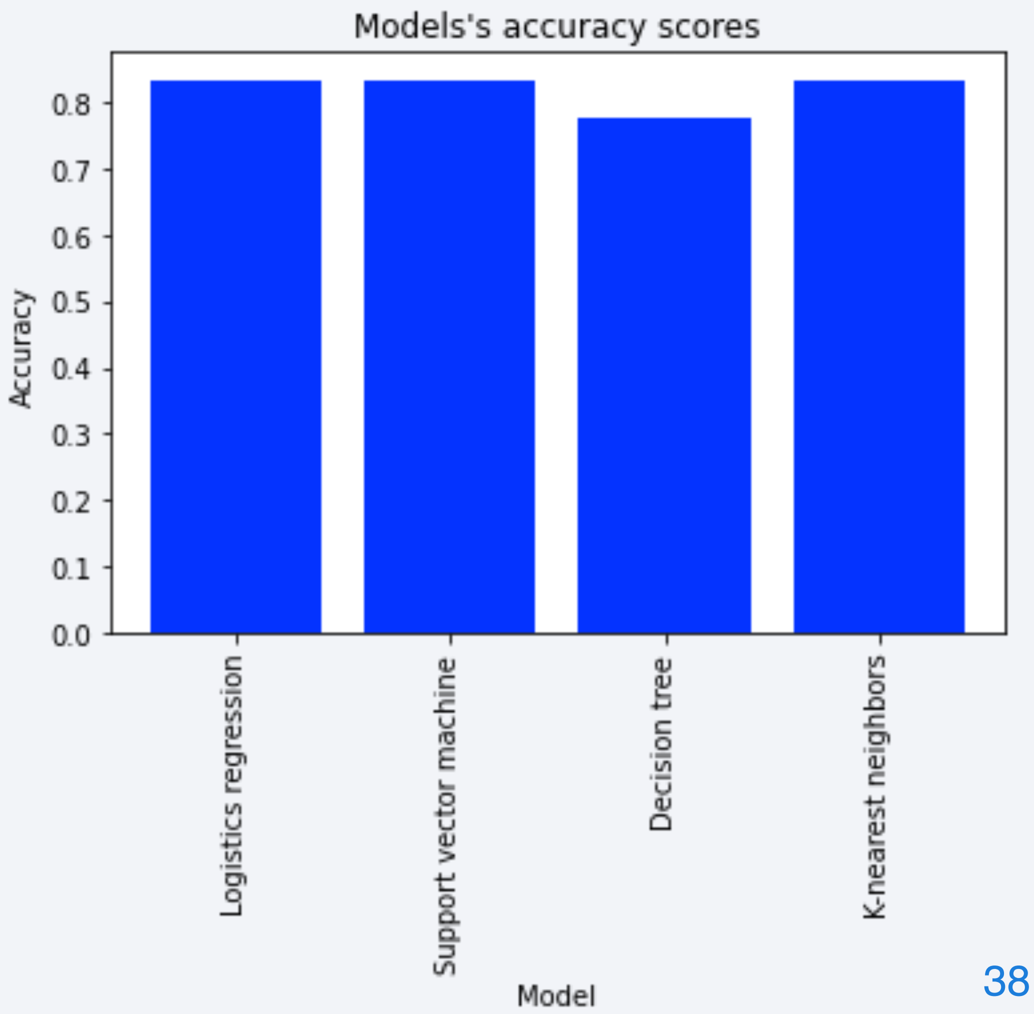

+
−
Ottawa
New York
Washington
United States
Phoenix
10
VAFB
Los Angeles
SLC-4E
46
The Bahamas
México
La Habana
Cuba
Ciudad
de México
República
Dominicana
Kingston
Ciudad
Honduras
Leaflet | Data by © OpenStreetMap, under ODbL.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Logistics regression, support vector machine and K-nearest neighbors all have the same accuracy of ~0.83



Confusion Matrix

- This confusion matrix looks the same for logistics regression, support vector machine and K-nearest neighbors
- Out of the 18 predictions, 15 has been predicted correctly.

Conclusions

- Low weight payloads deal better than high weight payloads
- KSC LC 39A had the most successful launches out of all the sites - although it could be connected with the fact that low weight payloads are used there
- Logistics regression, support vector machine and K-nearest neighbors all performed at ~0,83 accuracy

Appendix

- Notebooks can be found here:
  - https://github.com/0x1F0FFF/DS-ML-Capstone-Project

Thank you!