A Journey into Data Science

Evandro Oliveira

28-03-2023

EXECUTIVE SUMMARY

- Introduction.
- Data collection and data wrangling methodology.
- EDA and interactive visual analytics methodology.
- Predictive analysis methodology related.
- EDA with visualization results.
- EDA with SQL results.
- Interactive map with Folium results.
- Plotly Dash dashboard results.
- Predictive analysis (classification).
- Conclusion.

INTRODUCTION

- This is a complete review of the "Applied Data Science Capstone" course.
- The purpose of this presentation is to detail the various lessons learned during the course.
- We'll fly over all the steps of a practical application of data science.
- The theme of this work will be the prediction lading results of the Falcon-9 rocket missions.

Data collection and data wrangling methodology

- To collect data from a url the python requests library was used. We retrive a dataframe from json content from the SpaceX web API.
- The url contained information about SpaceX's falcon rocket missions. Relevant information was selected to analyze. Missing data were replaced by the mean value.

[74]:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitud
4	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.56185
5	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.56185
6	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.56185
7	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.63209
8	5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.56185

Data collection and data wrangling methodology

- Data was collected from Wikipedia Web regarding the launches of Falcon 9 and Falcon Heavy Launches.
- The retrieved HTML was parsed using the BeautifulSoup library.
- All necessary data has been extracted from within the html tags.

```
[63]: df=pd.DataFrame(launch_dict)
      df.head()
```

[63]:

	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	<a href="/wiki/Low_Earth_orbit" title="Low Ear...	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	<a href="/wiki/Low_Earth_orbit" title="Low Ear...	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	<a href="/wiki/Low_Earth_orbit" title="Low Ear...	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	<a href="/wiki/Low_Earth_orbit" title="Low Ear...	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	<a href="/wiki/Low_Earth_orbit" title="Low Ear...	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data collection and data wrangling methodology

- To perform an exploratory analysis of the data we calculated the number of launches at each location, the number and occurrence of each orbit, the number and occurrence of mission result by orbit type and finally created a landing outcome label from the column outcome.

[19]:
```
df.tail(5)
```

[19]:

LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
KSC LC 39A	True ASDS	2	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	2	B1060	-80.603956	28.608058	1
KSC LC 39A	True ASDS	3	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	2	B1058	-80.603956	28.608058	1
KSC LC 39A	True ASDS	6	True	True	True	5e9e3032383ecb6bb234e7ca	5.0	5	B1051	-80.603956	28.608058	1
CCAFS SLC 40	True ASDS	3	True	True	True	5e9e3033383ecbb9e534e7cc	5.0	2	B1060	-80.577366	28.561857	1
CCAFS SLC 40	True ASDS	1	True	False	True	5e9e3032383ecb6bb234e7ca	5.0	0	B1062	-80.577366	28.561857	1

EDA with SQL results

The following information was extracted from the SQL database:

- Launch sites are: CCAFS LC-40, VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40.
- The total payload lauched by Nasa was 99980 kg.
- The average payload mass carried by booster version F9 v1.1 was ~ 2534.67 kg.
- The first succesful landing outcome in ground pad was acheived in 22-12-2015 at 01:29h UTC time with Booster F9 FT B1019.
- The boosters which have success in drone ship and have payload mass greater than 4000 kg but less than 6000 kg are F9 FT B1022, F9 FT B1026, F9 FT B1021.2 and F9 FT B1021.2.
- The total number of successful and failure mission outcomes was 101.

EDA with SQL results

The following information was extracted from the SQL database:

- The names of the booster versions which have carried the maximum payload mass and the records which will display the month names, failure landing outcomes in drone ship in year 2015 are shown below:

```
[18]: %sql select distinct BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL)
 * sqlite:///my_data1.db
Done.
```

[18]:

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

```
[19]: %%sql select substr(Date, 4, 2) as Month,
    landing_outcome, Booster_Version, launch_site from SPACEXTBL
        where substr(Date,7,4)='2015' and Landing_Outcome like '%Failure (drone ship)%'
 * sqlite:///my_data1.db
Done.
```

[19]:

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

EDA with SQL results

The following information was extracted from the SQL database:

- The count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order was:

[47]:
```
%%sql select * from SPACEXTBL where Landing_Outcome
    like '%Suc%' and date(substr(Date,7,4)||'-'||substr(Date,4,2)||'-'||substr(Date,1,2))
        BETWEEN date('2010-06-04') AND  date('2017-03-20') order by Date DESC
```

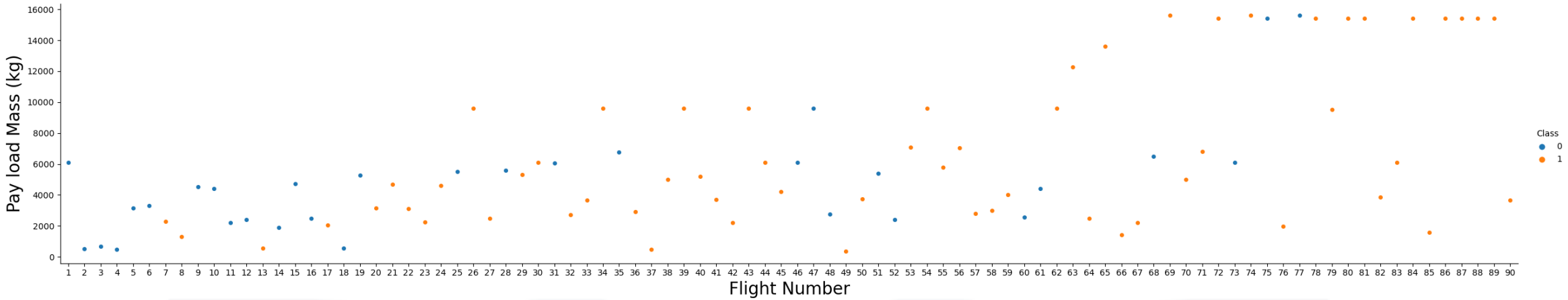
```
 * sqlite:///my_data1.db
Done.
```

[47]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
27-05-2016	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
22-12-2015	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)
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)
18-07-2016	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
14-08-2016	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
14-01-2017	17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
08-04-2016	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
06-05-2016	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)

EDA and interactive visual analytics methodology

- To perform an exploratory and visual analysis of the data, the matlibplot library was used. The following graph displays the relationship between flight number and payload:



EDA and interactive visual analytics methodology

- In order to prepare the data for the modeling step, we perform feature engineering we select the best features that could be used in success prediction. Next, we transform the columns into categorical data:

[57]:
```
features_one_hot = pd.get_dummies(features[['Orbit','LaunchSite','LandingPad','Serial']])
features_one_hot.head()
```

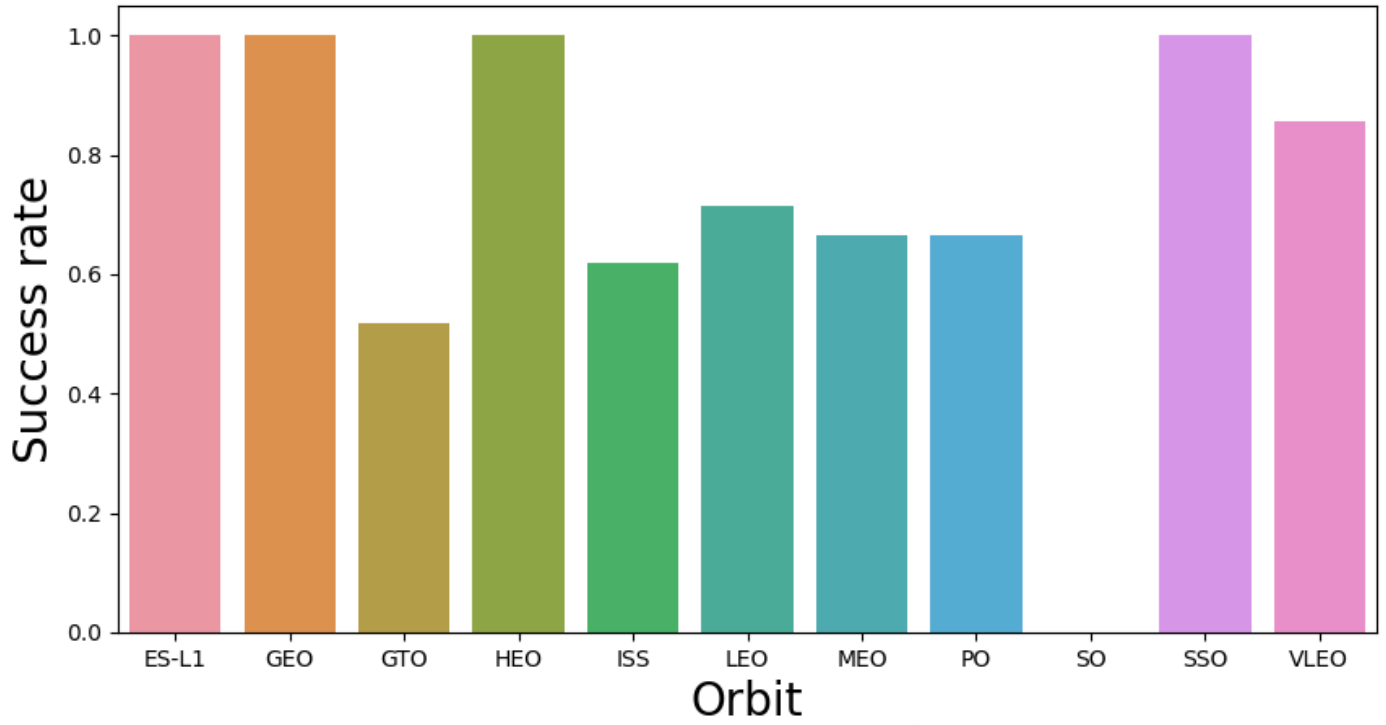
[57]:

	Orbit_ES-L1	Orbit_GEO	Orbit_GTO	Orbit_HEO	Orbit_ISS	Orbit_LEO	Orbit_MEO	Orbit_PO	Orbit_SO	Orbit_SSO	...	Serial_B1048	Serial_B1049	Serial_B1050	Serial_B1051	Se
0	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	
1	0	0	0	0	0	1	0	0	0	0	...	0	0	0	0	
2	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	
3	0	0	0	0	0	0	0	1	0	0	...	0	0	0	0	
4	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	

5 rows × 72 columns

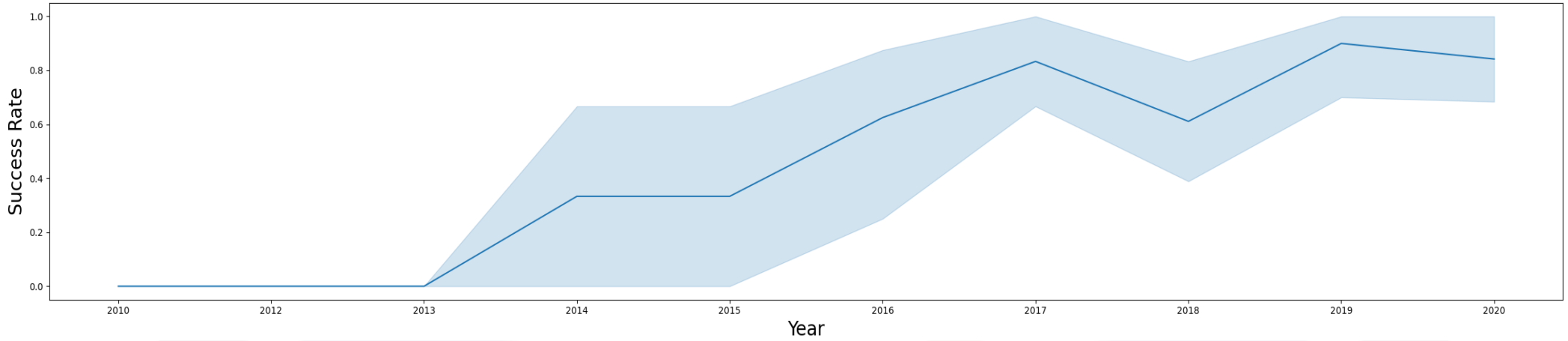
EDA with visualization results

- The following graph displays the relationship Between success rate and the type of Orbit:



EDA with visualization results

- The success of Falcon missions can be seen as a growing trend over the years:

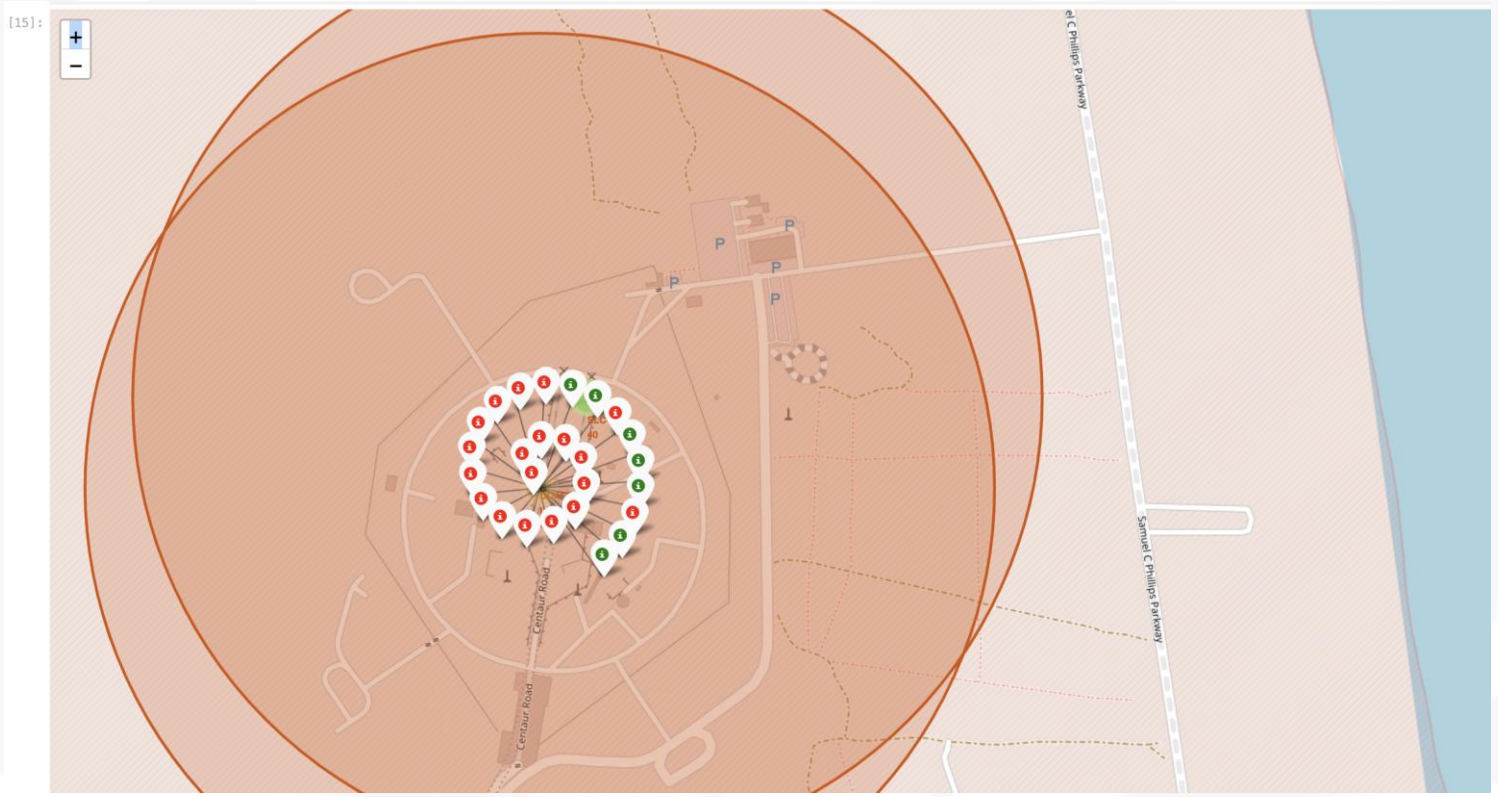


Interactive map with Folium results

- To better understand the launch data we performed an analysis of rocket launch sites using the folium library. First we added a tag for each US launch location:

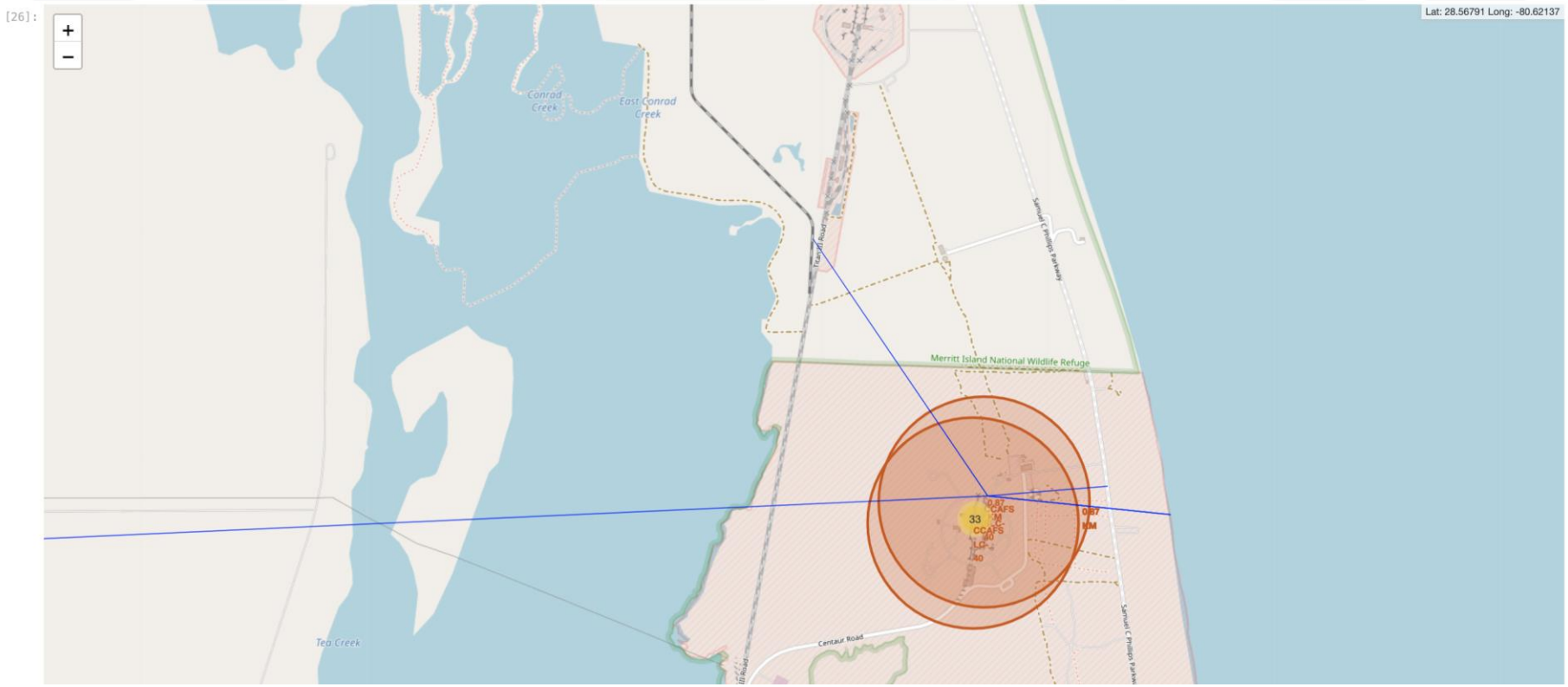
Interactive map with Folium results

- We've added a marker indicating each mission success or failure:



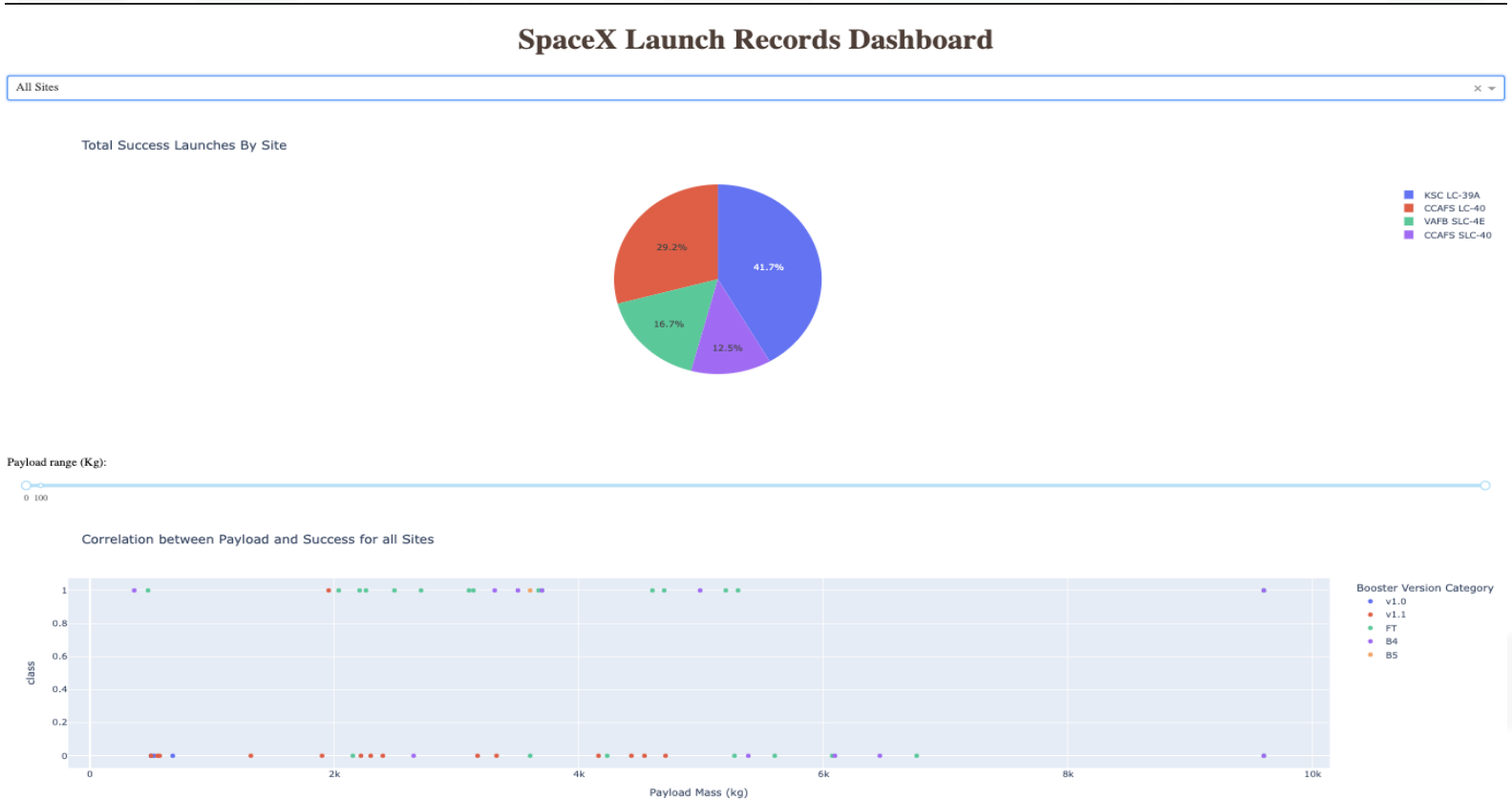
Interactive map with Folium results

- We calculate the distance to important points on the map showing the necessary safety distance for approval of orbital flights by the FAA.



Plotly Dash dashboard results

To get to know the data better we have prepared an interactive panel using the ploty dash library:



Plotly Dash dashboard results

From the created dashboard we answer the following questions:

- Which site has the largest successful launches? KSC LC-39A
- Which site has the highest launch success rate? KSC LC-39A
- Which payload range(s) has the highest launch success rate? Below 5500 kg.
- Which payload range(s) has the lowest launch success rate? Above 5500 kg.
- Which F9 Booster version (v1.0, v1.1, FT, B4, B5, etc.) has the highest launch success rate? B5

Predictive analysis methodology related

In order to predict the probability of a successful landing of the falcon 9 rocket, we adopted the strategy of testing some models and varying hyperparameters to find the best fit. We tested:

- Logistic Regression.
- Support Vector Machines.
- Decision Tree Classifier
- K-nearest Neighbors

Predictive analysis methodology related

Initially, we separate the Class column to serve as a label for the classification model:

```
[6]: data.head()
```

[6]:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

Second, we standardize the input features using a standardizer scaler and split the date into test dataset and validation dataset:

```
[93]: transform = preprocessing.StandardScaler()
      X =transform.fit_transform(X)
      X
```

```
[93]: array([[-1.71291154e+00, -1.94814463e-16, -6.53912840e-01, ...,
              -8.35531692e-01,  1.93309133e+00, -1.93309133e+00],
             [-1.67441914e+00, -1.19523159e+00, -6.53912840e-01, ...,
              -8.35531692e-01,  1.93309133e+00, -1.93309133e+00],
             [-1.63592675e+00, -1.16267307e+00, -6.53912840e-01, ...,
              -8.35531692e-01,  1.93309133e+00, -1.93309133e+00],
             ...,
             [ 1.63592675e+00,  1.99100483e+00,  3.49060516e+00, ...,
               1.19684269e+00, -5.17306132e-01,  5.17306132e-01],
             [ 1.67441914e+00,  1.99100483e+00,  1.00389436e+00, ...,
               1.19684269e+00, -5.17306132e-01,  5.17306132e-01],
             [ 1.71291154e+00, -5.19213966e-01, -6.53912840e-01, ...,
              -8.35531692e-01, -5.17306132e-01,  5.17306132e-01]])
```

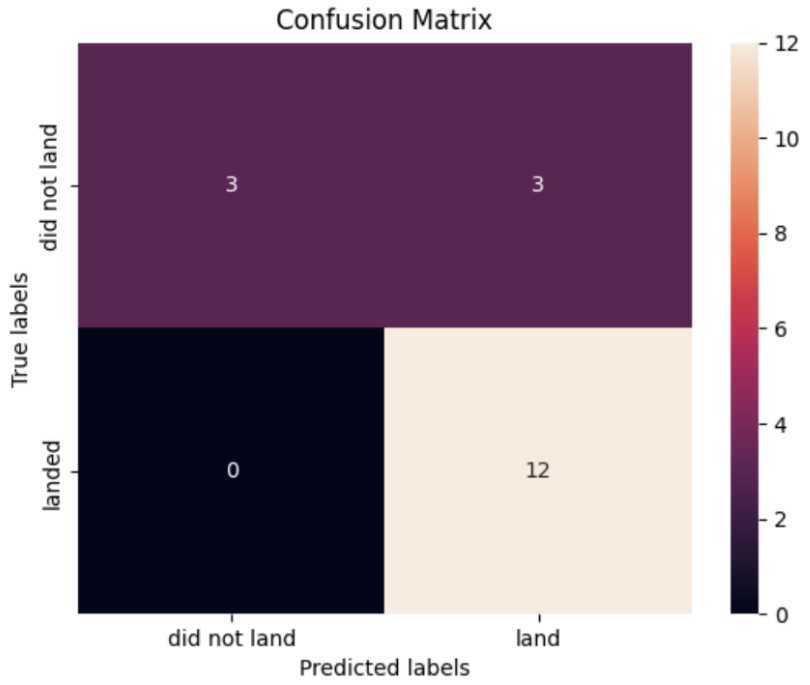
Predictive analysis (classification)

After training the models in calculating the accuracy based on the test dataset:

- Logistic Regression accuracy: 0.8333.
- Support Vector Machines accuracy: 0.833.
- Decision Tree Classifier accuracy: 0.778.
- K-nearest Neighbors accuracy: 0.833.

Predictive analysis (classification)

We also plot the confusion Matrix for each model, the major problema was the false positives. The confusion matrix was the same for all tested models.



Conclusion

- Acording the results the models Logistic Regression, Support Vector Machines accuracy and K-nearest Neighbors accuracy had the same accuracy results.
- Any one of this models coud be used for classification.
- Ther perfomance of the models was not so high.
- We recomended collect more data and select more features.

Insights

To improve the accuracy of the model, telemetry information from the model could be incorporated into the training data like:

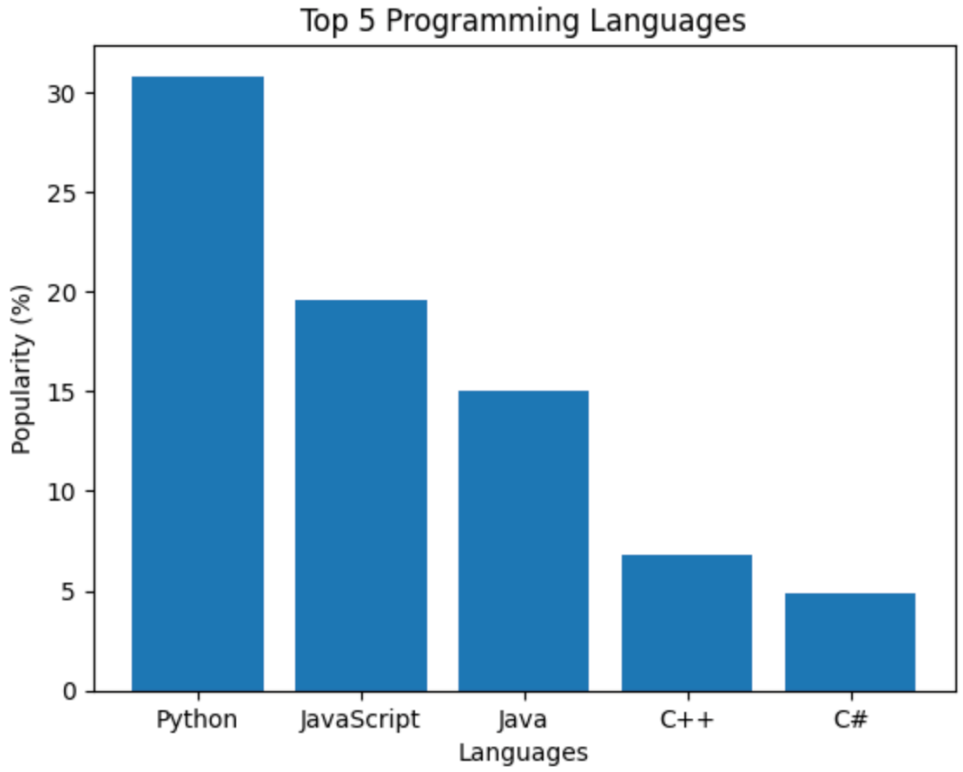
- Takeoff speed.
- Time to orbit.
- Angle of re-entry.
- Max-Q.
- Maximum speed.

APPENDIX

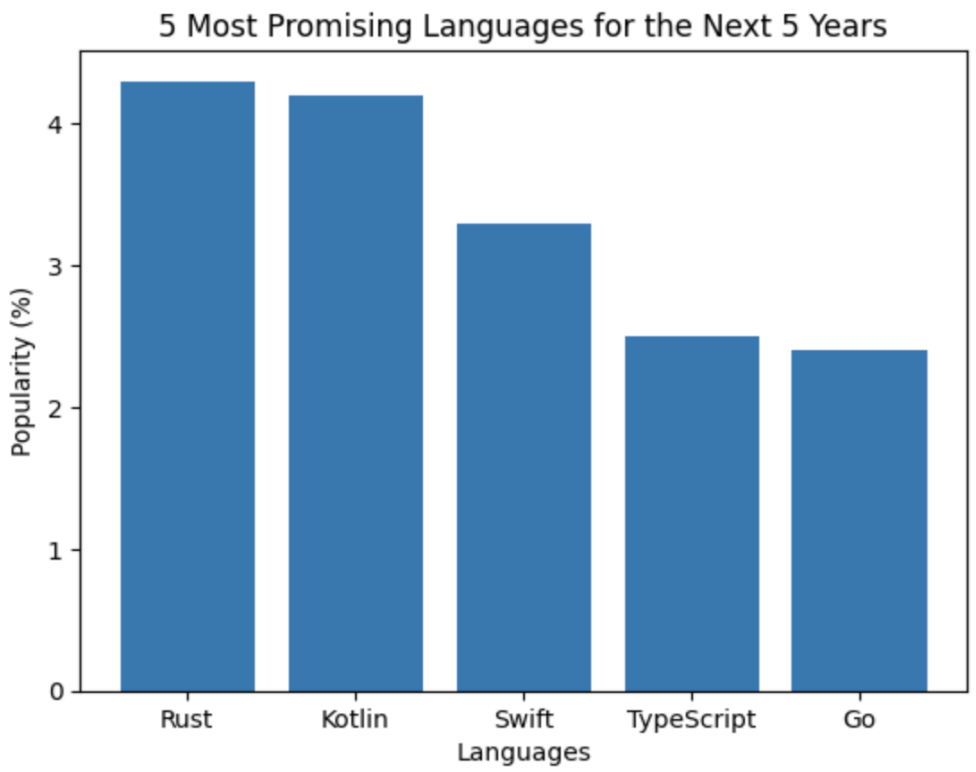
- Programming language trends.
- Popular languages from kaggle dataset.
- Programming language findings & implications.
- Database Trends.
- Database Trends & Implications.
- Flight count by airline to destination state.

PROGRAMMING LANGUAGE TRENDS

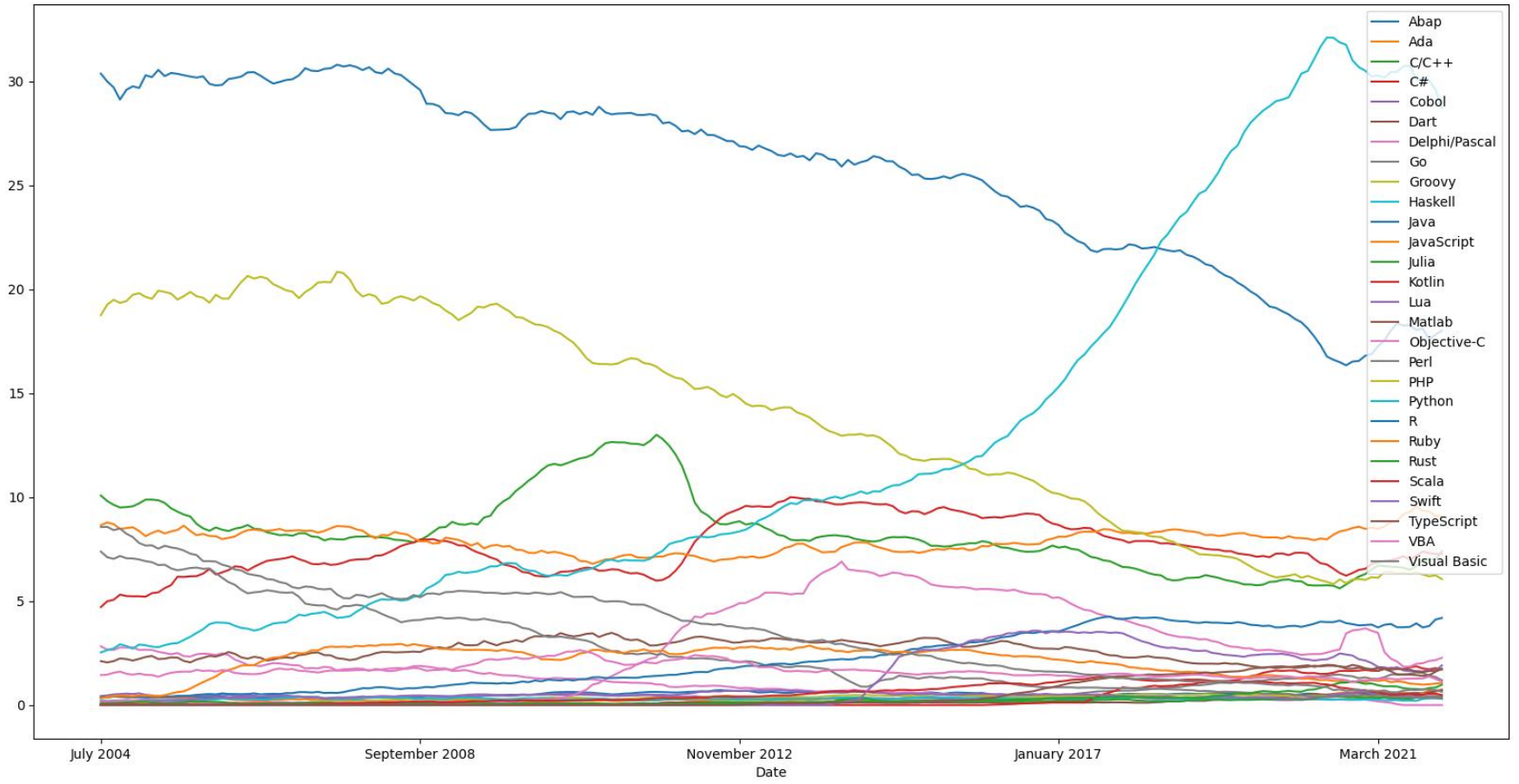
Current Year



Next Year



POPULAR LANGUAGES



PROGRAMMING LANGUAGE TRENDS - FINDINGS & IMPLICATIONS

Findings

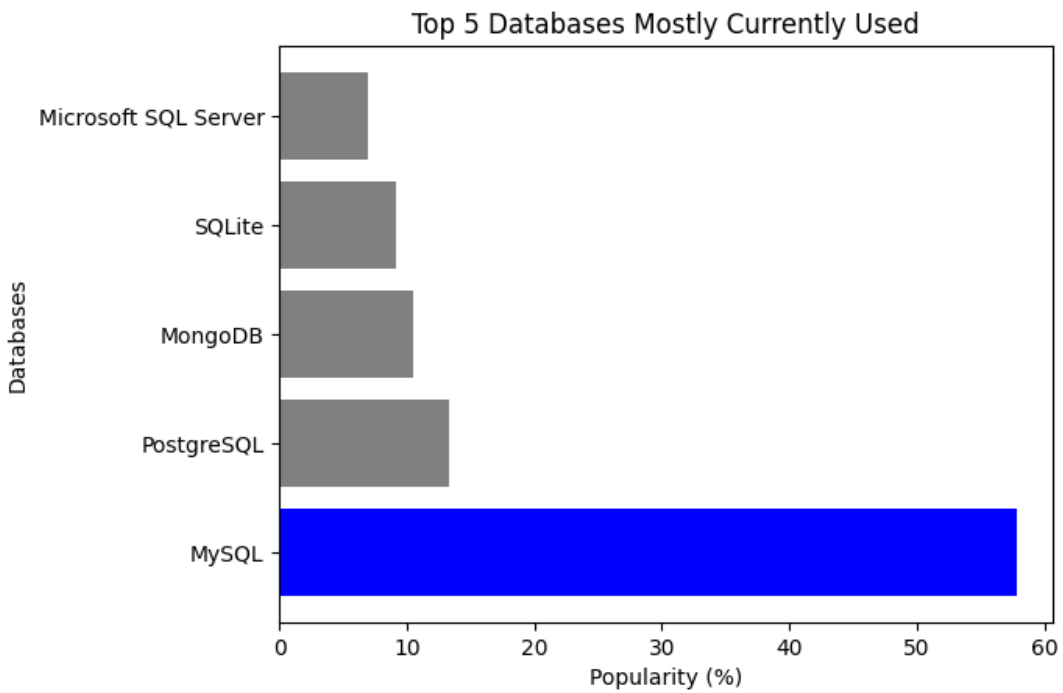
- Python is the most used language for data science in the world today.
- Javascript is the most used language on the web.
- C++ continues to have an important relevance.

Implications

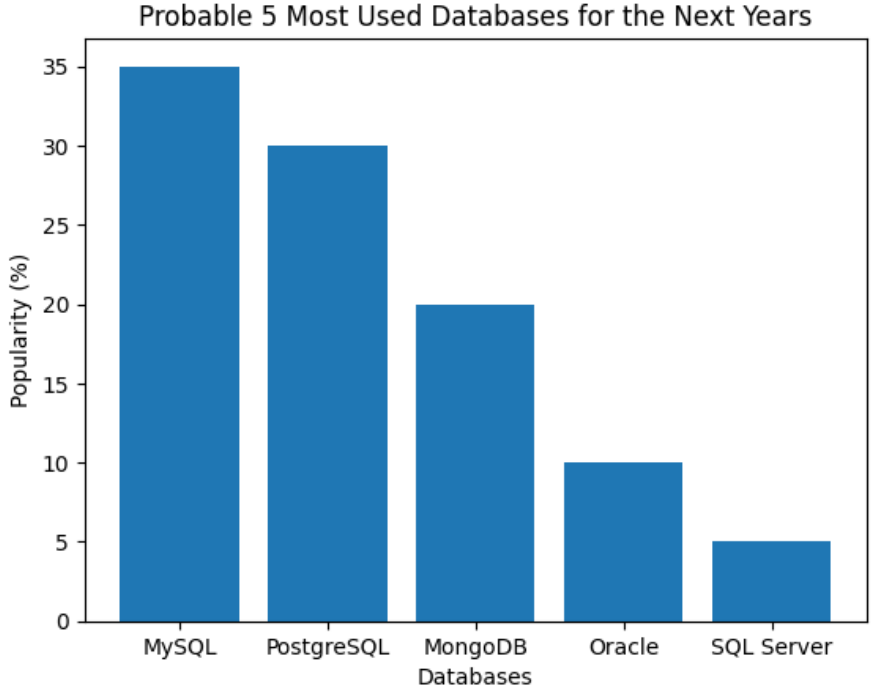
- Rust should gain great relevance in network solutions.
- Go could gain many applications in machine learning.
- Typescript should continue to gain many fans.

DATABASE TRENDS

Current Year



Next Year



DATABASE TRENDS - FINDINGS & IMPLICATIONS

Findings

- MySQL is the most used database in the world.
- Postgres remains an extremely viable option.
- MongoDB well represents the non-relational banks with a good portion of the market.

Implications

- SQL Server will continue to be the most used database.
- Oracle databases seem to be on a trend to regain lost market share.

Flight count by airline to destination state