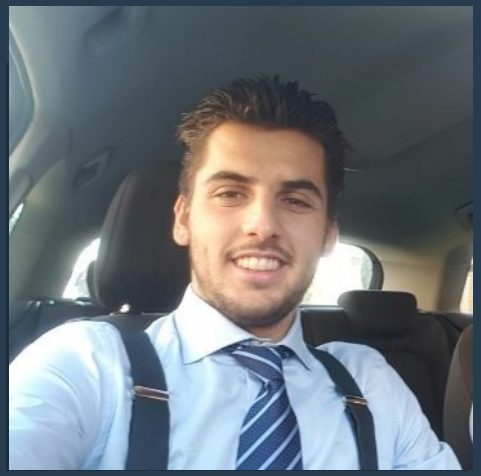
CROWDFUNDING

DATA PREPARATION AND ANALYSIS

Meet The Crowdfunding Team

Cristina Blanco

Group Member



Alvaro Gericke

Group Member

Marta Larrea

Group Member

Tien Tran

Group Member

Presentation Agenda

Introduction and Project Presentation

EDA and Feature Selection

Data Sources and Merging

Modeling

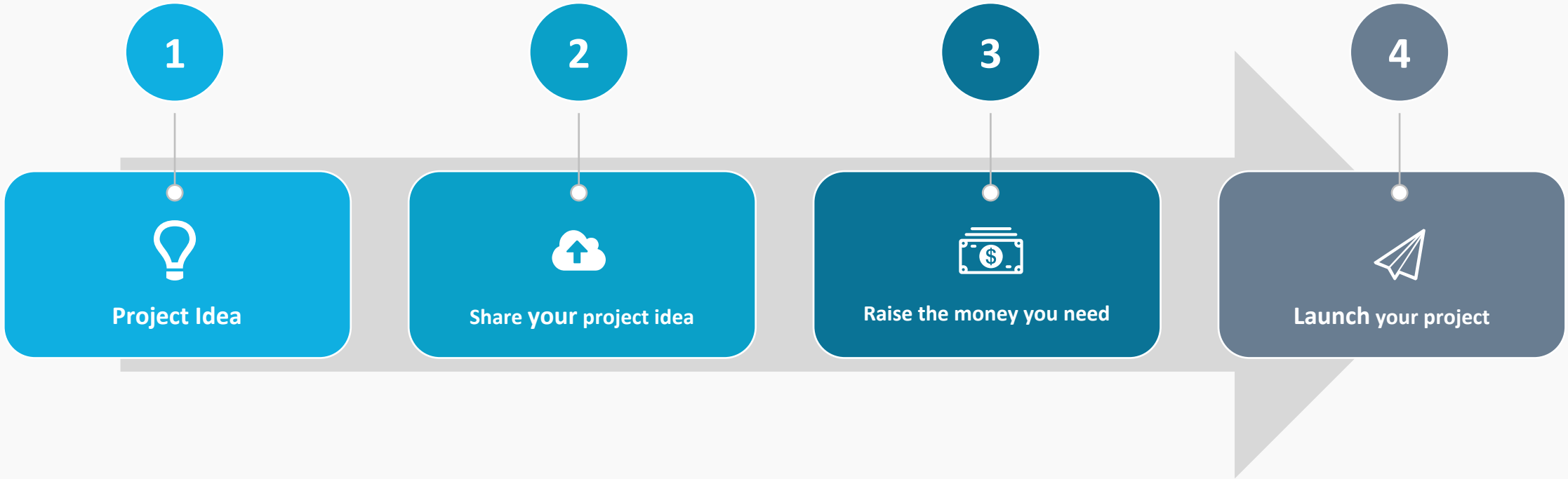
Data Cleaning and Feature Engineering

Conclusions

Introduction & Project Description

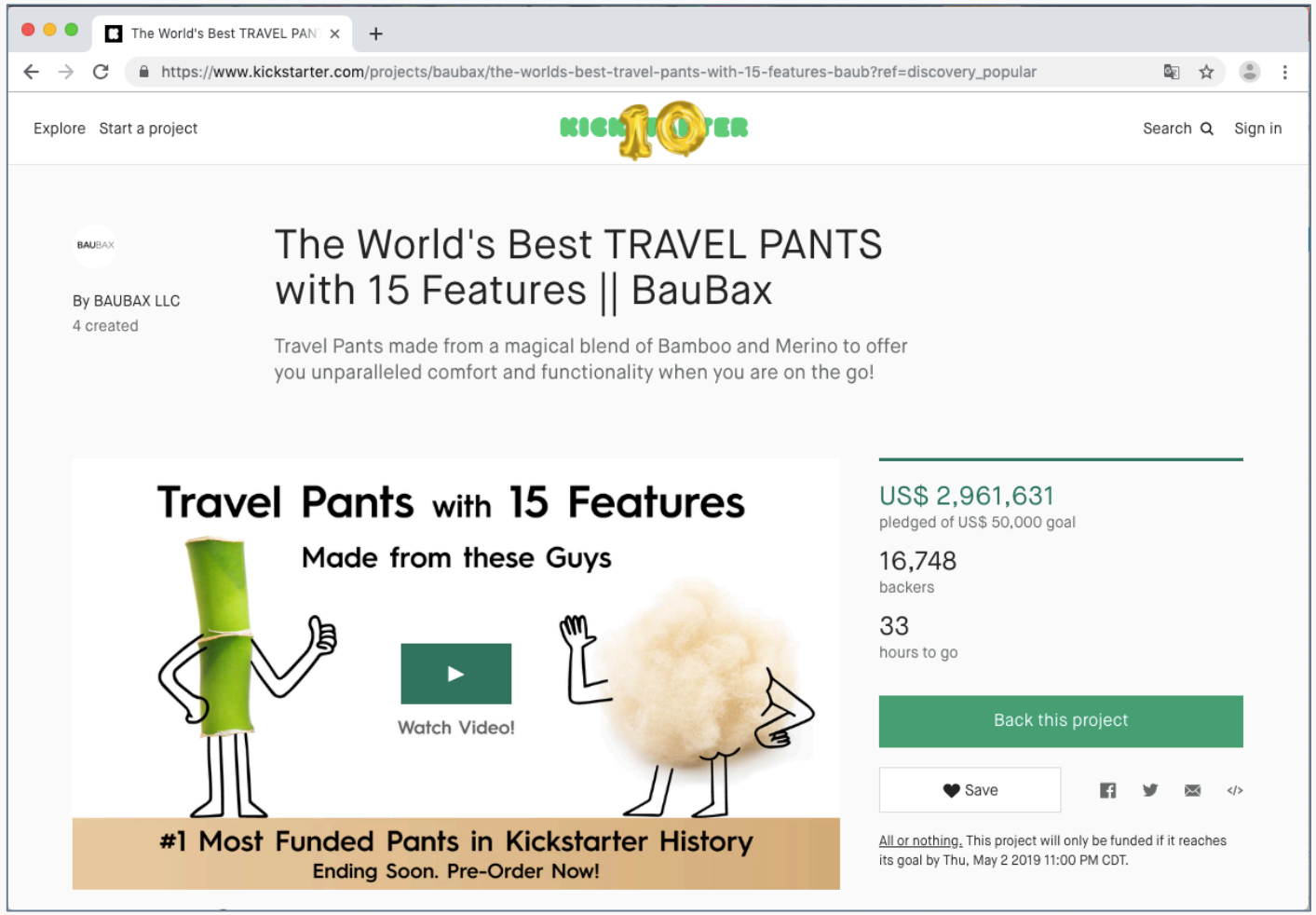
What Is Crowdfunding?

The practice of funding a project or venture by raising many small amounts of money from a large number of people



Kickstarter Platform

Kickstarter helps artists, musicians, filmmakers, designers, and other creators find the resources and support they need to make their ideas a reality.



Platform Stats

To date, tens of thousands of creative projects — big and small — have come to life with the support of the Kickstarter community.

$4,264,617,953
Total dollars pledged to Kickstarter projects

443,362
Launched projects

162,391
Successfully funded projects

16,199,092
Total backers

Amount of money to raise

If I ask for a lot of money, will my project fail?

Launching Date

What is the best day to launch my proposal?

Location

Which country has the highest successful rate?

Duration of the Campaign

How long should the collection last?

Category

Does my project belong to a popular category?

What is our goal?

With this project we want to get a general and clear view of which projects would be more successful than others, taking into account characteristics such as the goal money, the duration of the campaign, the region in which it is launched and some others. We also want to determine the success rate, that a certain project would have, in order to help future entrepreneurs decide if launching their project is worth it or not.

Data Sources & Merging

Where did we get our data?

We get it from **WebRobots.io**

They have a scraper robot which crawls all Kickstarter projects and collects data in CSV and JSON formats. From March 2016 they run this data crawl once a month.

- We downloaded the .csv files using a web scrapping script
- Afterwards, we merged all files in a single data set.
- We have only to download the files from 2016 to 2018.

Dataset Description

Variable	Description
Id	Each crowdfunding campaign has an id
Photo	It contains the photo information
Name	Name of the project
Blurb	Description of the project
Goal	The amount of money to raise
Pledged	The amount of money actually raised
State	Categorical variable with the values: failed, canceled, live, successful and suspended
Slug	Another name for the project

Variable	Description
Disable_communication	Logical variable with false true. Disable communication with backers
Country	Country were the crowdfunding is started
Currency	The currency in which the money is being raised
Currency_symbol	Denotes the symbol of the currency in which the money is being raised
Currency_trailing_code	Logical variable with false true
Deadline	The date till they have to have raised the money
State_changed_at	The date in which the state variable changed
Created_at	The date in which the crowdfunding project was created

Variable	Description
Launched_at	The date in which the crowdfunding project was launched
Backers_count	Amount of sponsor.
Staff_pick.	Logical variable with false true, that determines if it was highlighted by the staff
Static_usd_rate	Conversion rate in US dollars
Usd_pledged	Conversion in US dollars of the pledged column
Creator	Contains the creators user url information
Location	Contains the url information of the country and city were the crowd funding was launched
Category	Dictionary containing the url information of the category of the project

Data Cleaning & Feature Engineering

Dataset Cleaning

Variable	Use
Id	Unique variable used for merging all .csv files
Photo	No use
Name	Used to create a numerical value. The length in words of the name
Blurb	Used to create a numerical value. The length in words of the description
Goal	Used for modelling and analysis
Pledged	Used to create other variables for modelling
State	Target variable 
Slug	No use

Variable	Use
Disable_communication	No use
Country	Use only as support, because many of the variables are wrongly attributed
Currency	Used for modelling and analysis
Currency_symbol	No use
Currency_trailing_code	No use
Deadline	Variable used to create the duration variable
State_changed_at	Used when merging the .csv to obtain the most recent update
Created_at	Used for analysis

Variable	Use
Launched_at	Used to create variables such as the year, month and weekday in which it was launched, and the duration
Backers_count	Used for modeling and analysis
Staff_pick.	Not used because there were many missing values
Static_usd_rate	Used to convert the goal into usd_goal
Usd_pledged	Used to create the success rate variable
Creator	No use
Location	Used to obtain the correct country and other variables
Category	Used to obtain the main category of the project.

Missing Values

Missing Values per Column: **Name <** 0.001% **Blurb <** 0.005% **Location ≈** 0.34 %

Distribution of missing values grouped by Location

Country	Amount
United States	1087
Great Britain	2
Denmark	1
Austria	1

Distribution of the missing values gruped by Category

Category	Amount
art	118
comics	14
crafts	9
dance	18
design	12
fashion	6
film & video	279
theater	22

Category	Amount
food	10
games	49
journalism	55
music	258
photography	49
publishing	141
technology	51

Feature engineering

Final selection of variables used in the Explanatory Data Analysis

Variable	Extraction Process
backers_count	Original variable
country	Original variable
currency	Original variable
id	Original variable
name	Original variable
state	Original variable
usd_pledged	Original variable
usd_goal	Goal variable in us dollars, calculated using the Static_usd_rate
duration	Since the project is launched until its due date

Feature engineering

Final selection of variables used in the Explanatory Data Analysis

Variable	Extraction Process
days_until_launched	Since the project is created until it is launched
year_launched	Extracted from the launched_at variable
month_launched	Extracted from the launched_at variable
weekday_launched	Extracted from the launched_at variable
name_length	The length in words of the name variable
description_length	The length in words of the blurb variable
main_category	From the category dictionary
sub_category	From the category dictionary
type	From the location dictionary

Feature engineering

Final selection of variables used in the Explanatory Data Analysis

Variable	Extraction Process
region_state	From the location dictionary
country2	From the location dictionary
pledge_per_backer	Calculated by dividing the usd_pledged and the backers_count
success_rate	Calculated by dividing the usd_pledged and the usd_goal
goal_range	Intervals are selected based on the statistics web page of kickstarter
goal_cat_division	Goal range group by categories
competitors	Group by category, goal_range, year and month
comp_range	From the competitors
competitors_cat_division	Competitors group by category

Missing Values

Missing Values per Column: **region_state** <0.0274% **pledge_per_backer** ≈11.335% **sub_category**≈ 9.28 %

Distribution of the missing values gruped by Category

main_category	Sub_category count	Sub_category percentage
art	2454	9.99
comics	2535	25.18
crafts	1901	28.78
dance	1634	50.20
design	959	4.53
fashion	2280	11.65
film&video	1072	2.48
food	3220	17.75

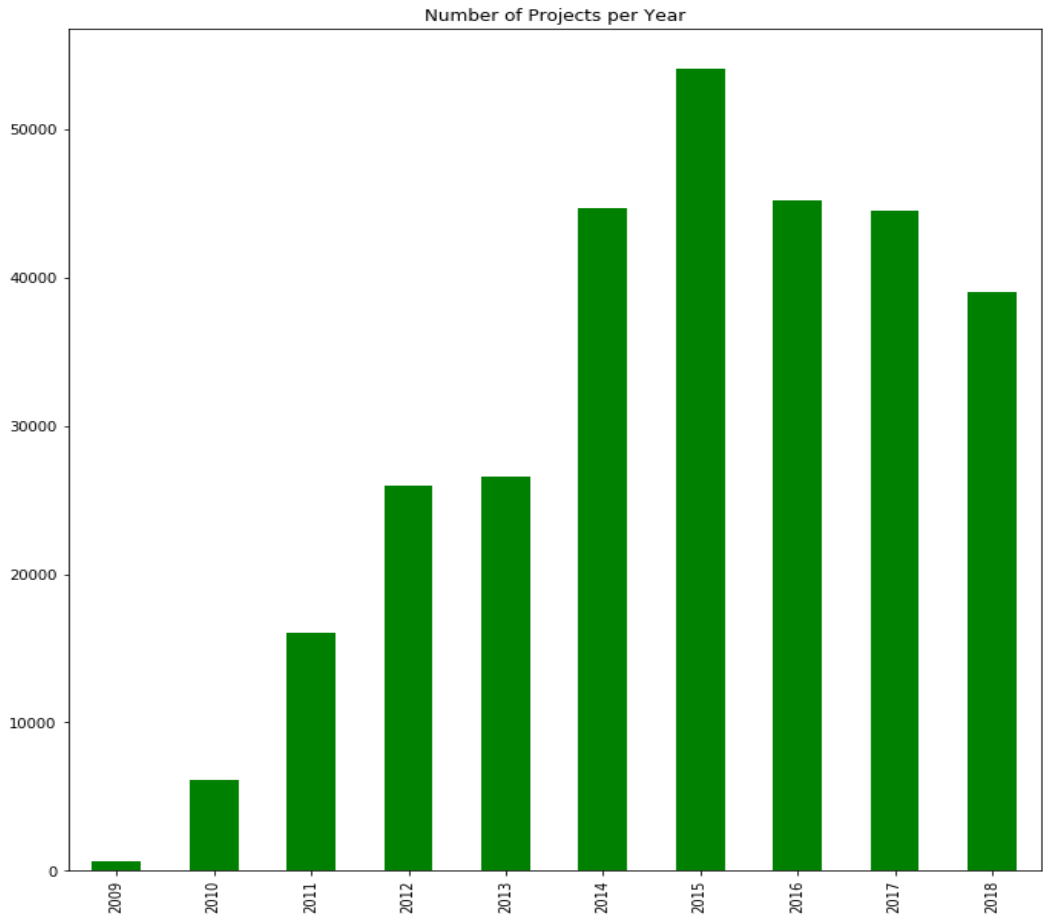
main_category	Sub_category count	Sub_category percentage
games	619	1.98
journalism	1461	30.29
music	2346	5.91
photography	1937	25.65
publishing	1170	3.31
technology	1584	5.23
theater	2961	40.29

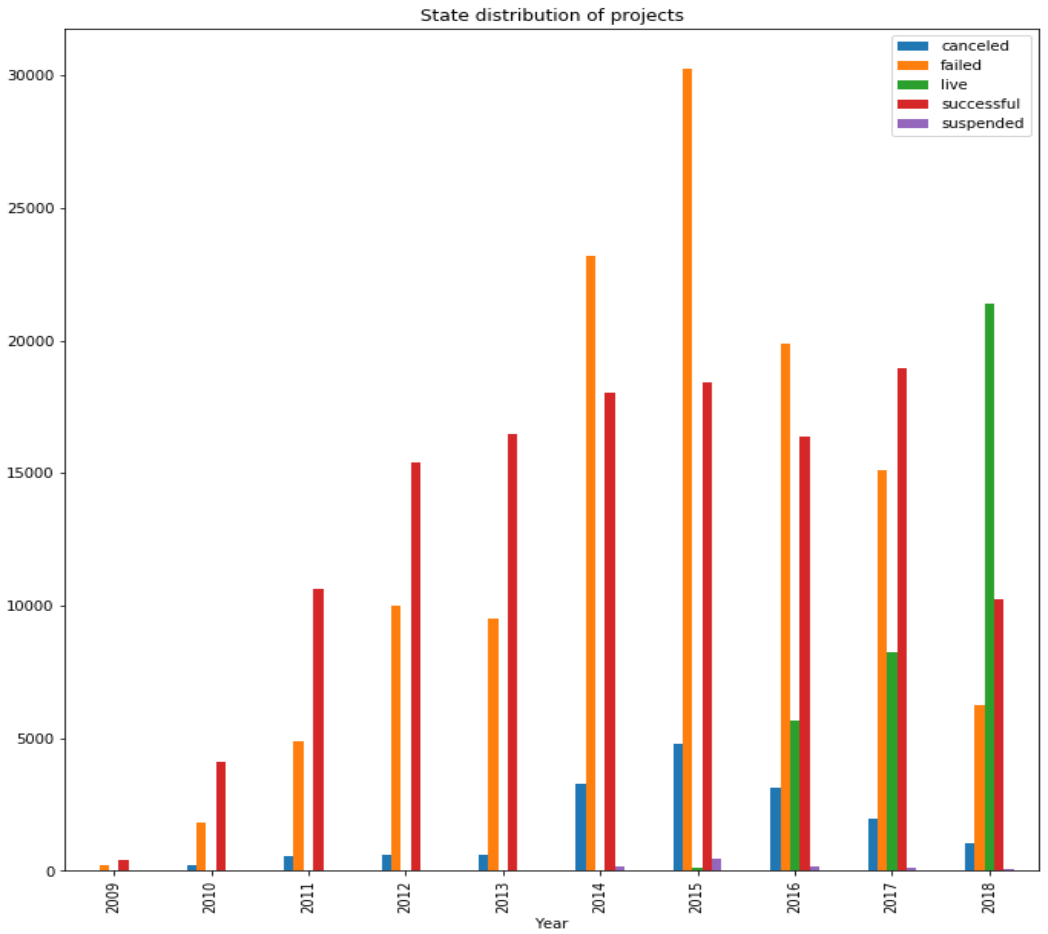
Distribution of missing values grouped by Country

Country	Region_state count
AQ Antartica	23
NZ New Zealand	23
MK Macedonia	15
XK Kosovo	7
GI Gibraltar	4
CW Curacau	3
SX Sint Maarten	3
KI Kiribati	1
VA Vatican City	1
AW Aruba	1
MO Macao	1
PN Pitcairn	1

EDA &
Feature Selection

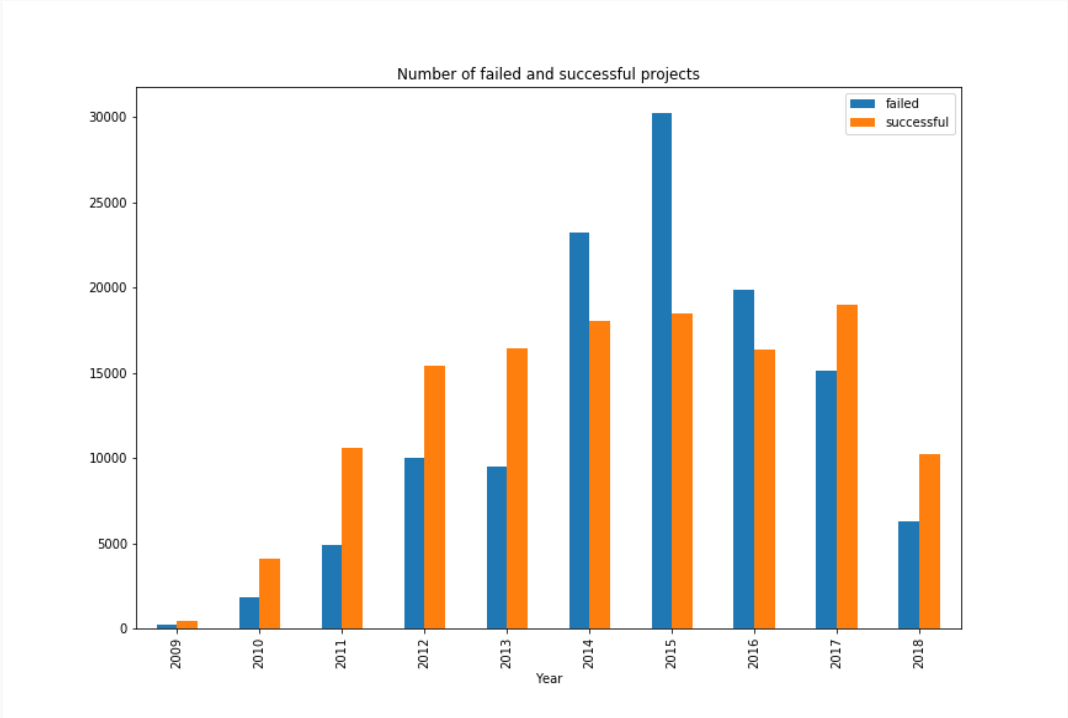
General Overview

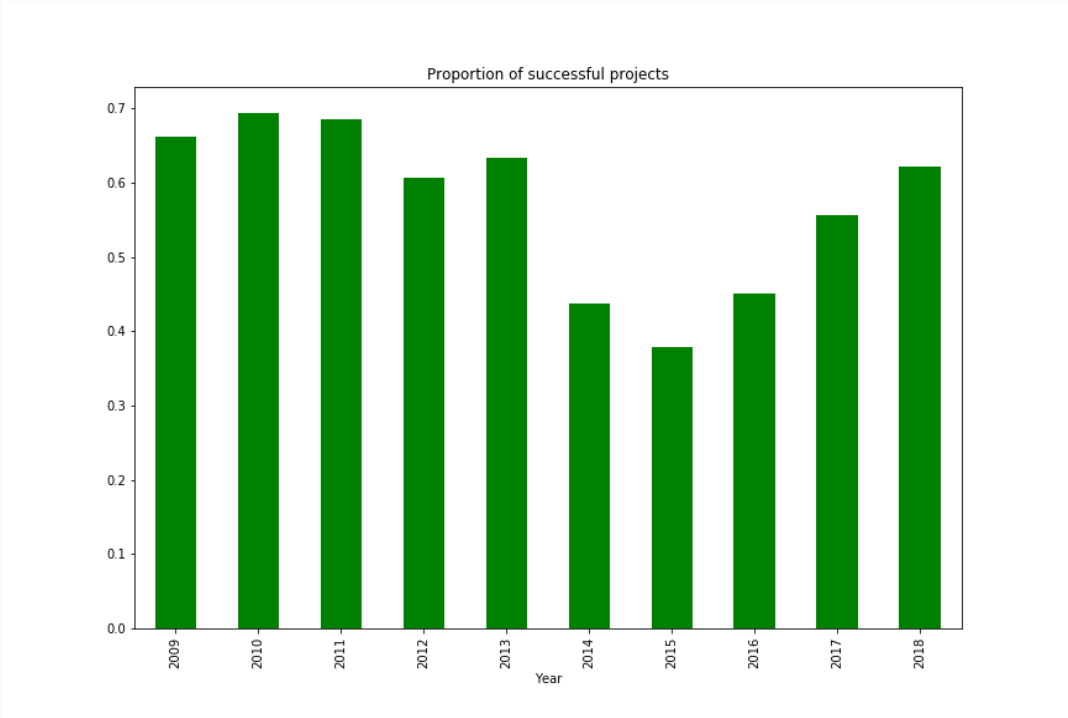




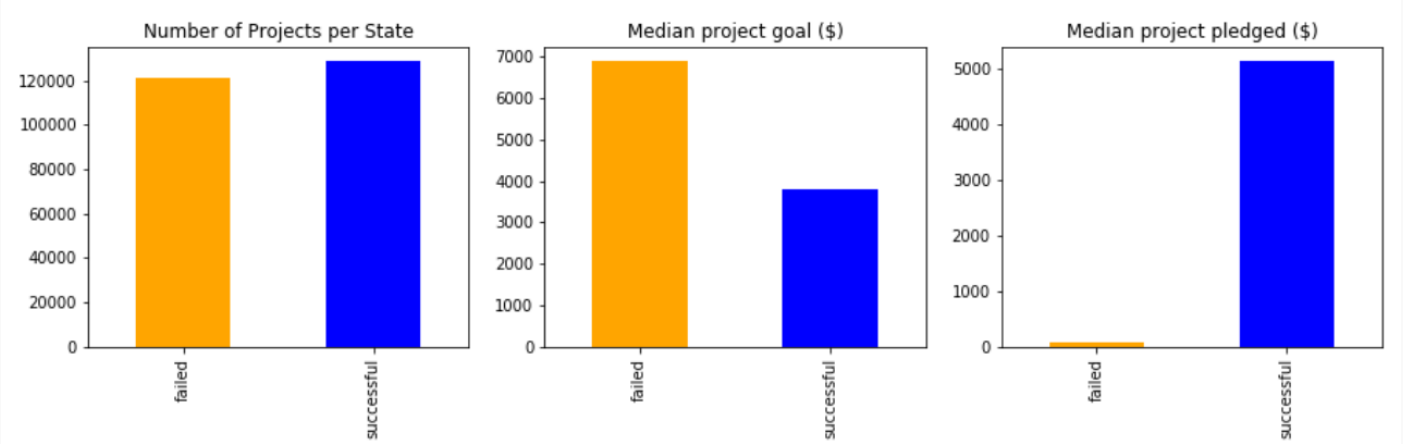
State Data Distribution

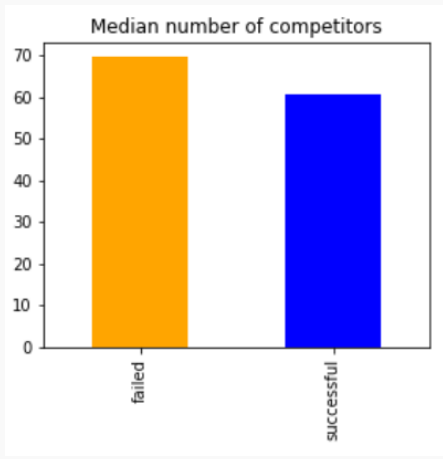
Successful	Failed	Live	Canceled	Suspended
42.62%	39.99%	11.70%	5.34%	0.34%

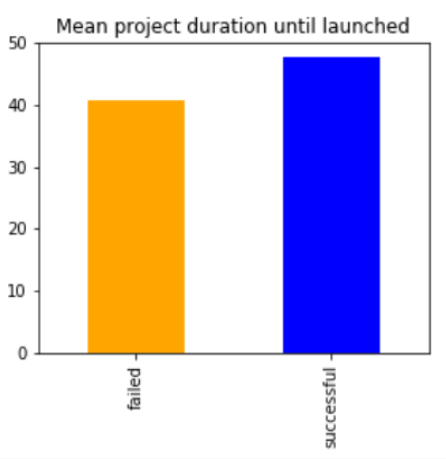


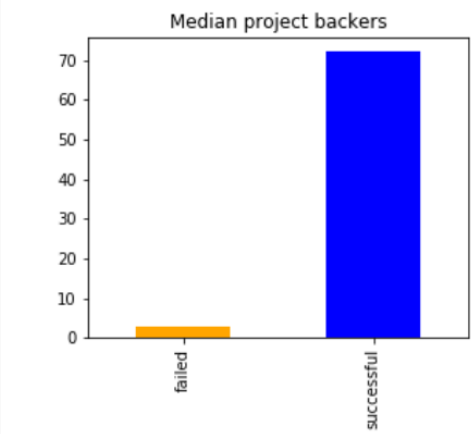


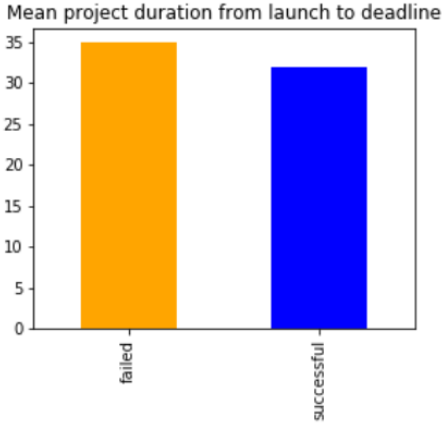
Failed VS Successful

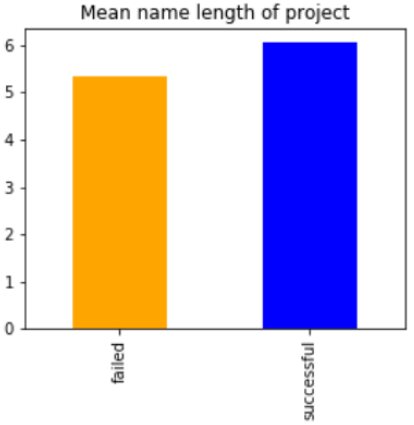


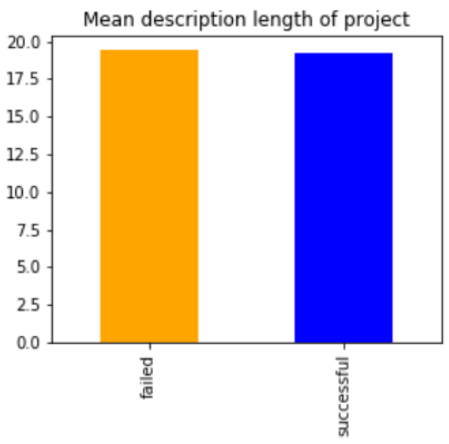




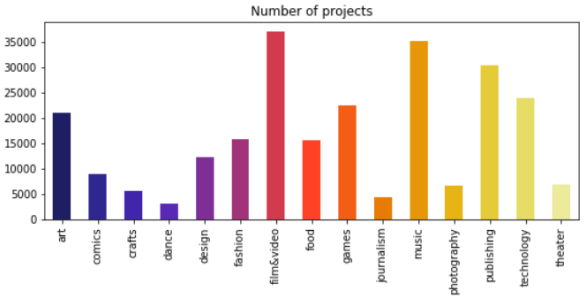


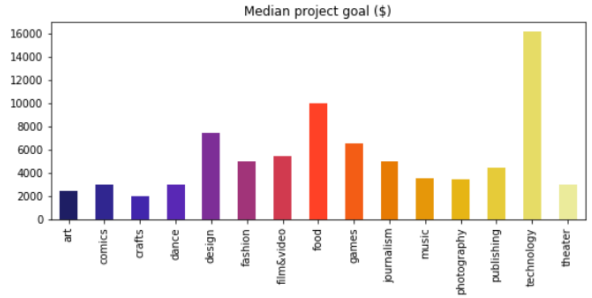


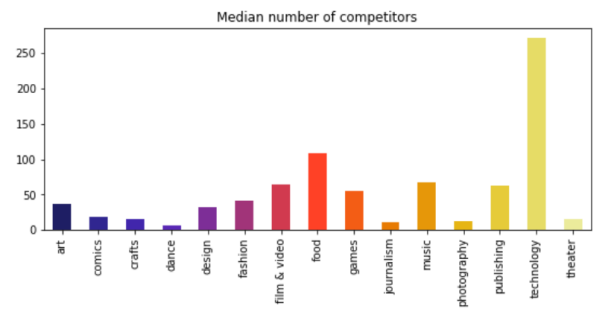


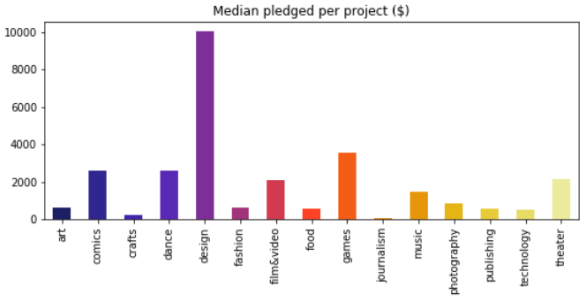


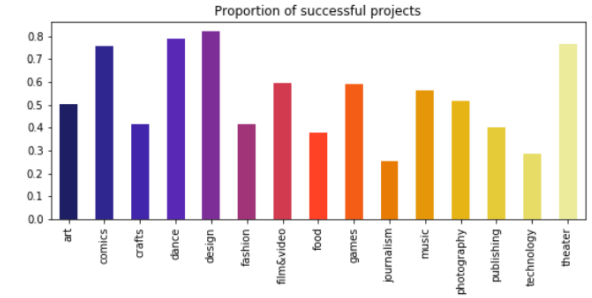
Categories Analysis

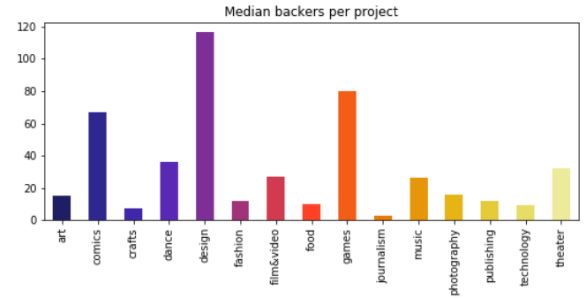


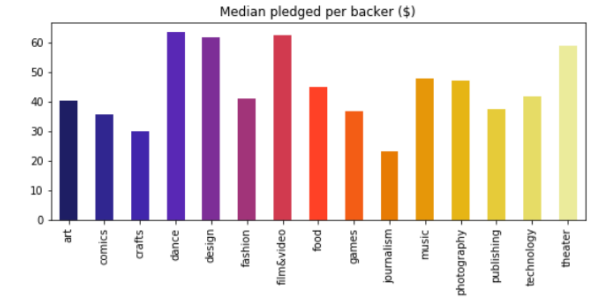








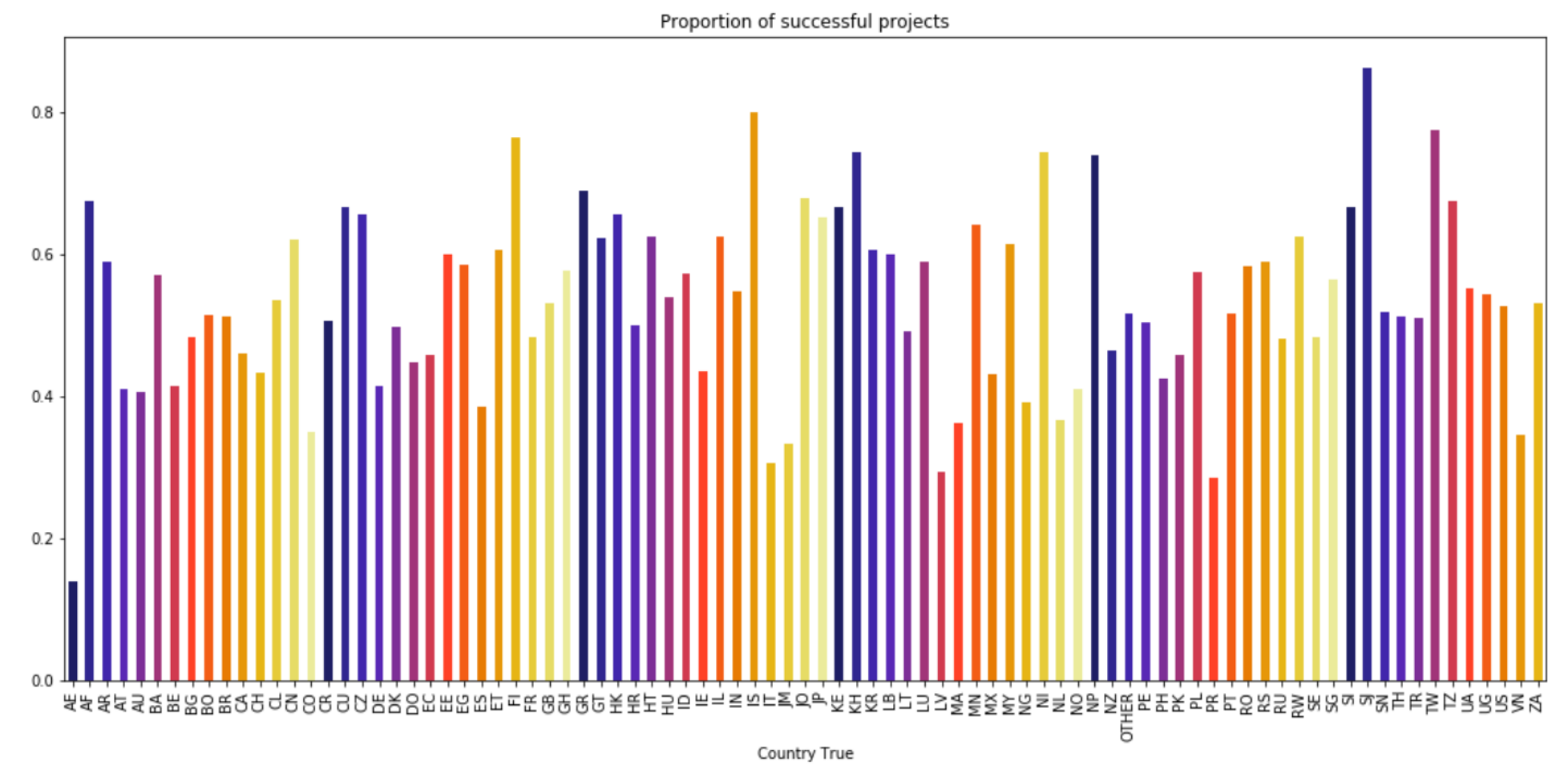




- Best: Design, dance, comic and theater.
- Worst: Technology and journalism.

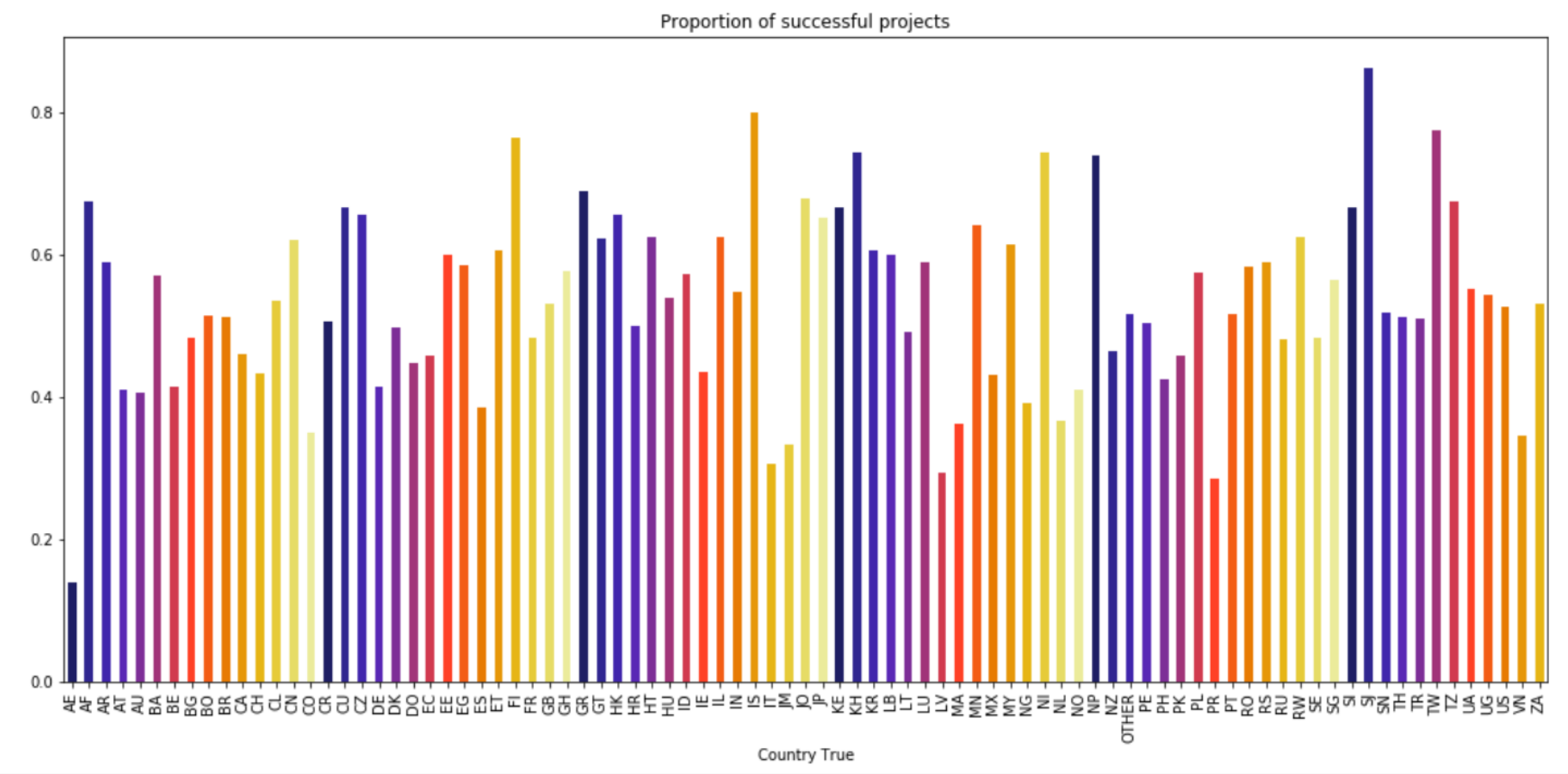
Country Analysis

Best: Islandia, Taiwan, Finlandia, Cambodia, Nicaragua or Nepal Worst: Arab Emirates, Puerto Rico, Latvia, Italy, Jamaica, Vietnam or Morocco

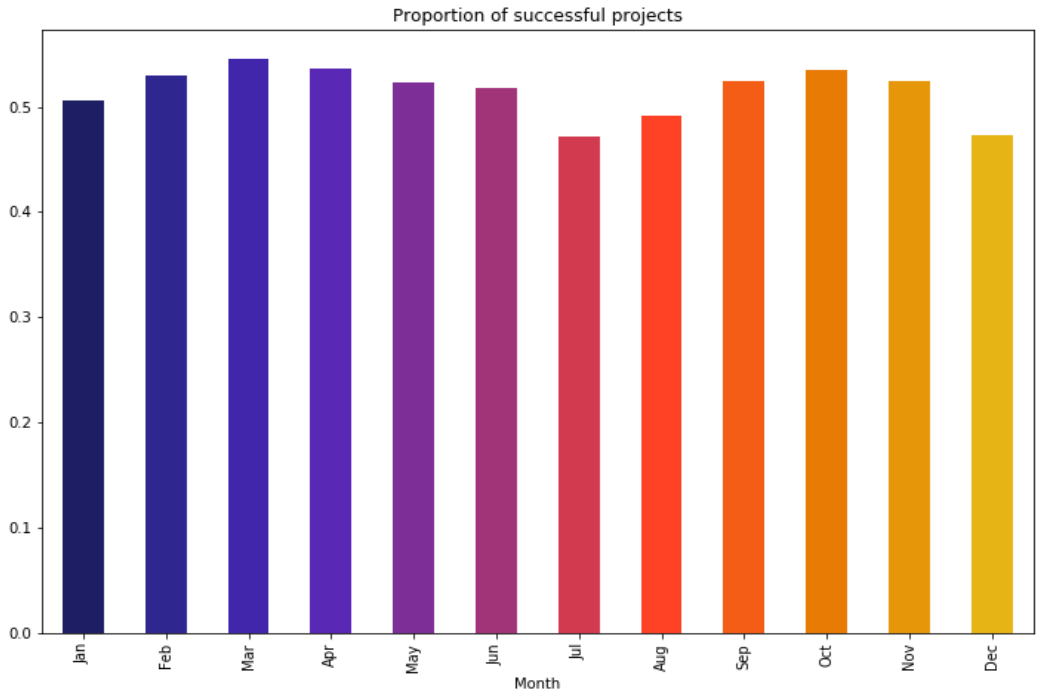


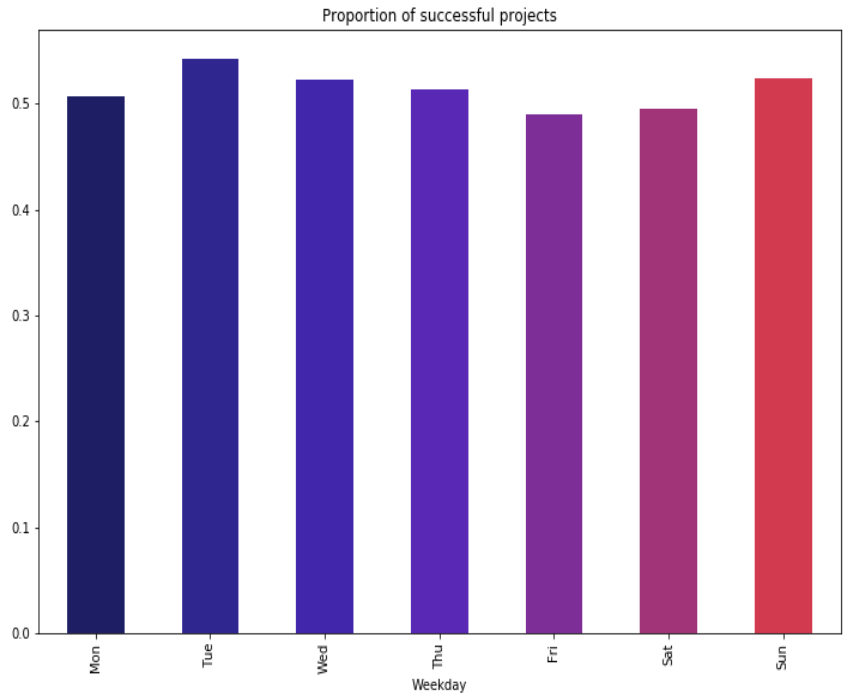
US State Analysis

Best: Vermont, New York, Massachusetts, Rhode Island, Oregon, Washington, Minnesota or California Worst: Florida, Mississippi, South Dakota, Arkansas, Alabama and Kansas

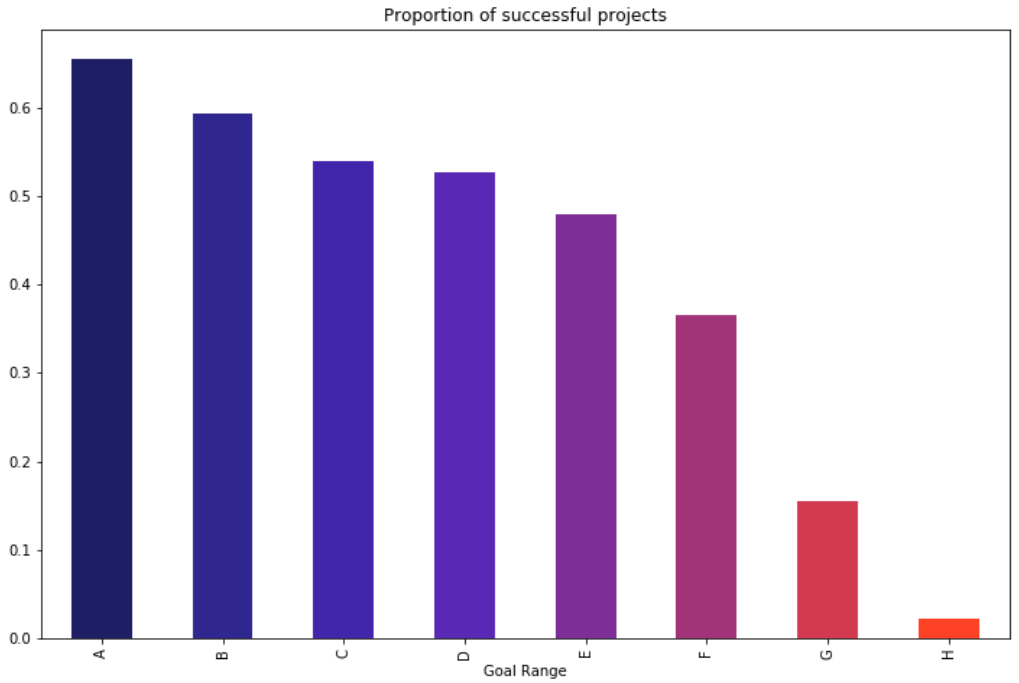


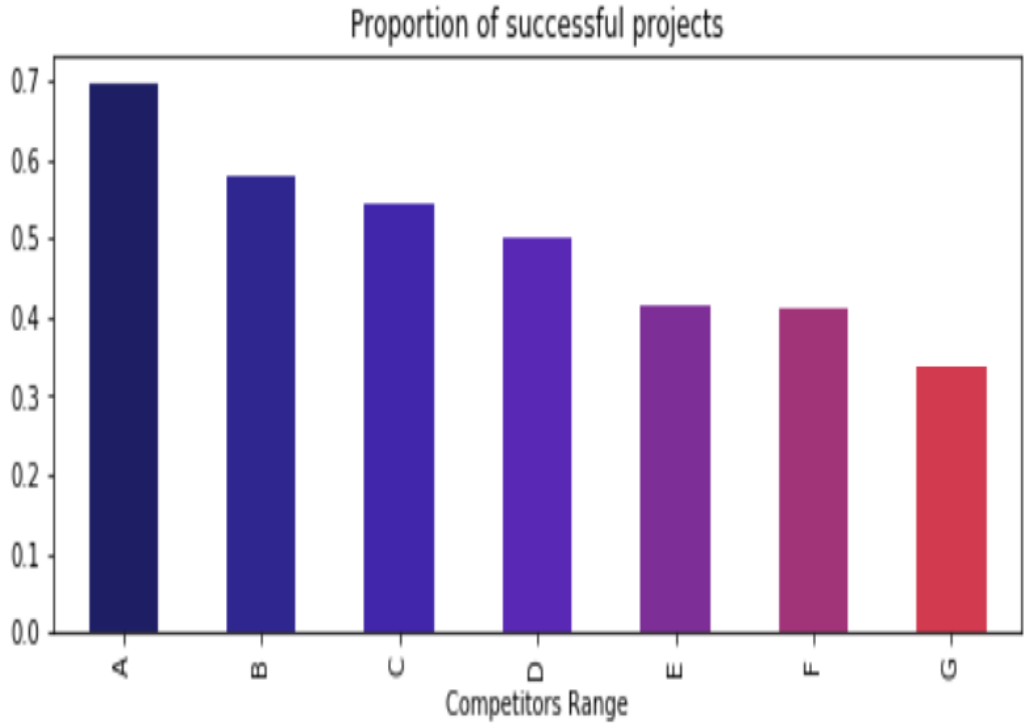
Month and Weekday Analysis





Goal and Competitors Analysis





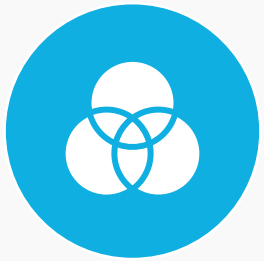
Modeling

Modeling Features and Targets

Features	
country2	description_length
currency	main_category
usd_goal	sub_category
duration	type
days_until_launched	goal_range
year_launched	goal_cat_division
month_launched	competitors
weekday_launched	comp_range
name_length	competitors_cat_division

Target
state
success_rate

Classification Models Description

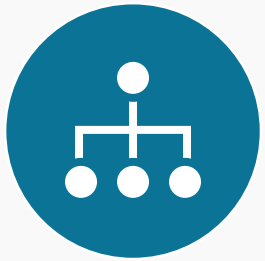


Naïve Bayes

We have decided to use the Gaussian Naïve Bayes classifier

Logistic Regression

We used the default parameters. We use recursive feature selection



Decision Tree

min_samples_split = 2 (easier to make a split)
min_samples_leaf = 1 (smaller leaf nodes)
max_depth = None (tree height is not restricted)

Random Forest

The tree growing criteria is the same as in the Decision Tree, and the default number of trees in the Random Forest is 10 (which is quite small).

GridSearchCV

After fitting the default decision tree and default random forest. We used grid search with cross validation to do the hyper parameter tuning

Decision Tree

Parameter Grid:

- 'max_depth': [None, 5, 10, 15, 20]
- 'min_samples_leaf': [1, 2, 3, 5]
- 'min_samples_split': [2, 5, 10]

Best Hyperparameters:

- 'max_depth': None
- 'min_samples_leaf': 1
- 'min_samples_split': 2

Random Forest

Parameter Grid:

- 'n_estimators': [10, 50, 100, 500]
- 'max_depth': [None, 5, 10, 15, 20]
- 'min_samples_leaf': [1, 2, 3, 5]
- 'min_samples_split': [2, 5, 10]

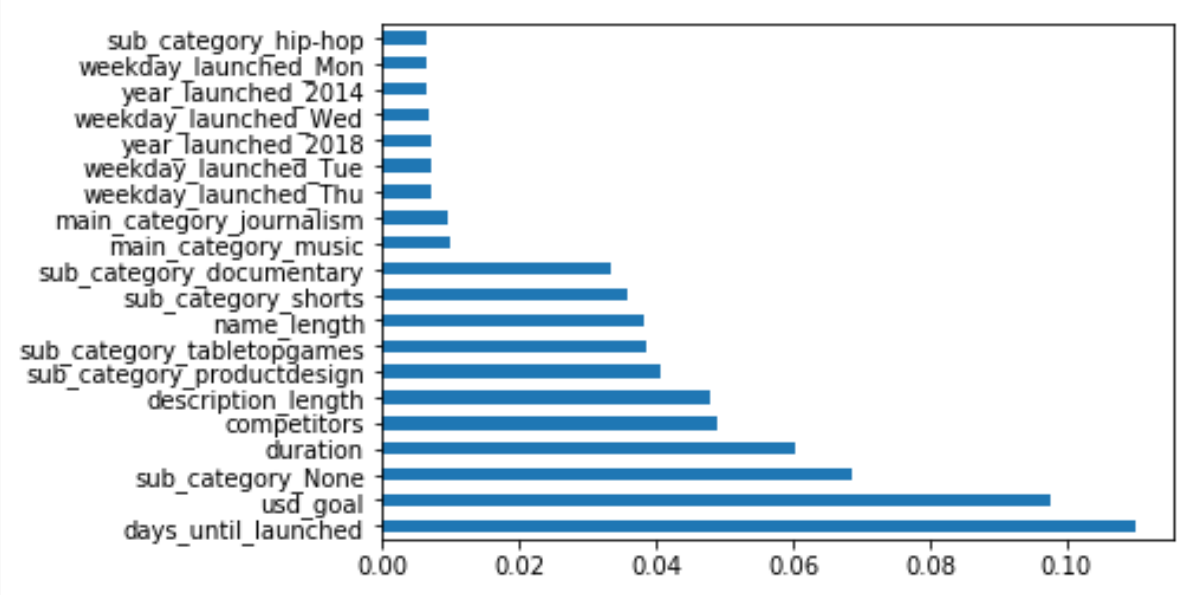
Best Hyperparameters:

- 'n_estimators': 500
- 'max_depth': None
- 'min_samples_leaf': 1
- 'min_samples_split': 2

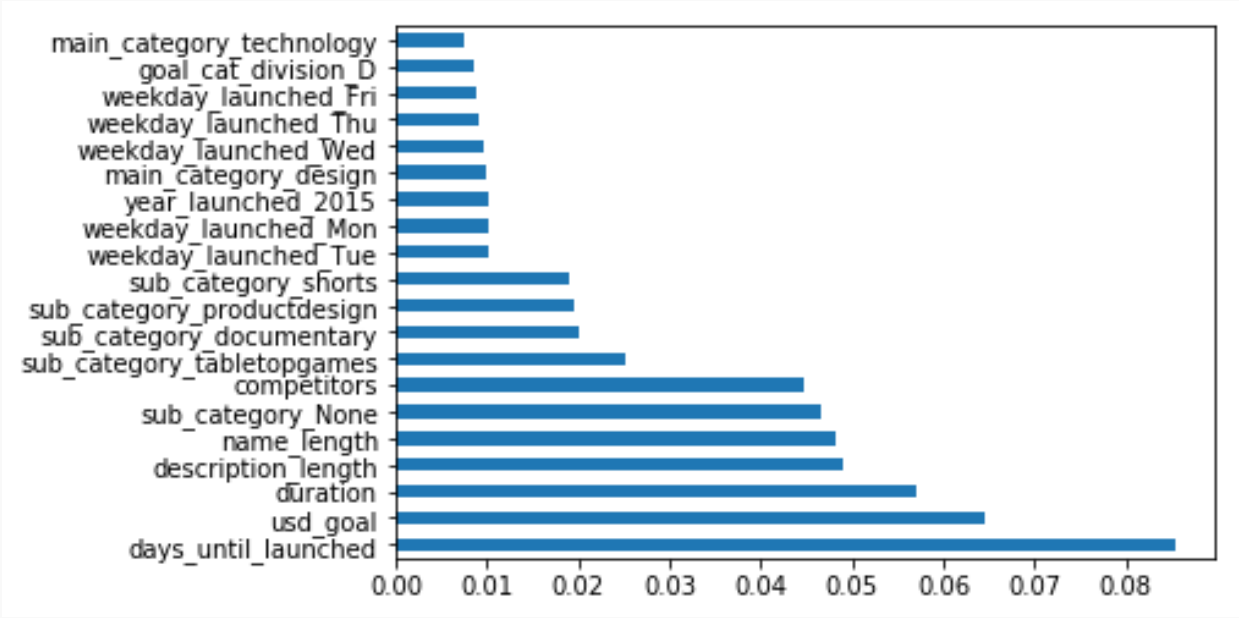
We use a 5-fold cross validation to see which parameters in the grid works best with the data set

Feature Importance

Decision Tree



Ramdon Forest

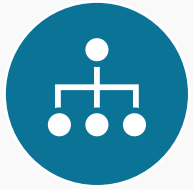


Classification Models Summary

	Accuracy	Precision	Recall
Naive Bayes	54.16	52.90	98.86
Logistic Regression	58.96	57.23	79.94
Logistic Regression w/ feature selection	58.96	57.23	79.94
Decision Tree	70.78	71.56	71.63
Decision Tree GridSearchCV	74.53	80.03	67.24
Random Forest	74.46	78.83	68.81
Random Forest GridSearchCV	77.89	80.75	68.02

Every classification model are validated using a hold-out test set of 20% the total data

Regression Models



Decision Tree

Similar to the decision tree classifier, the parameter supports growing a tree as large as possible.

Random Forest

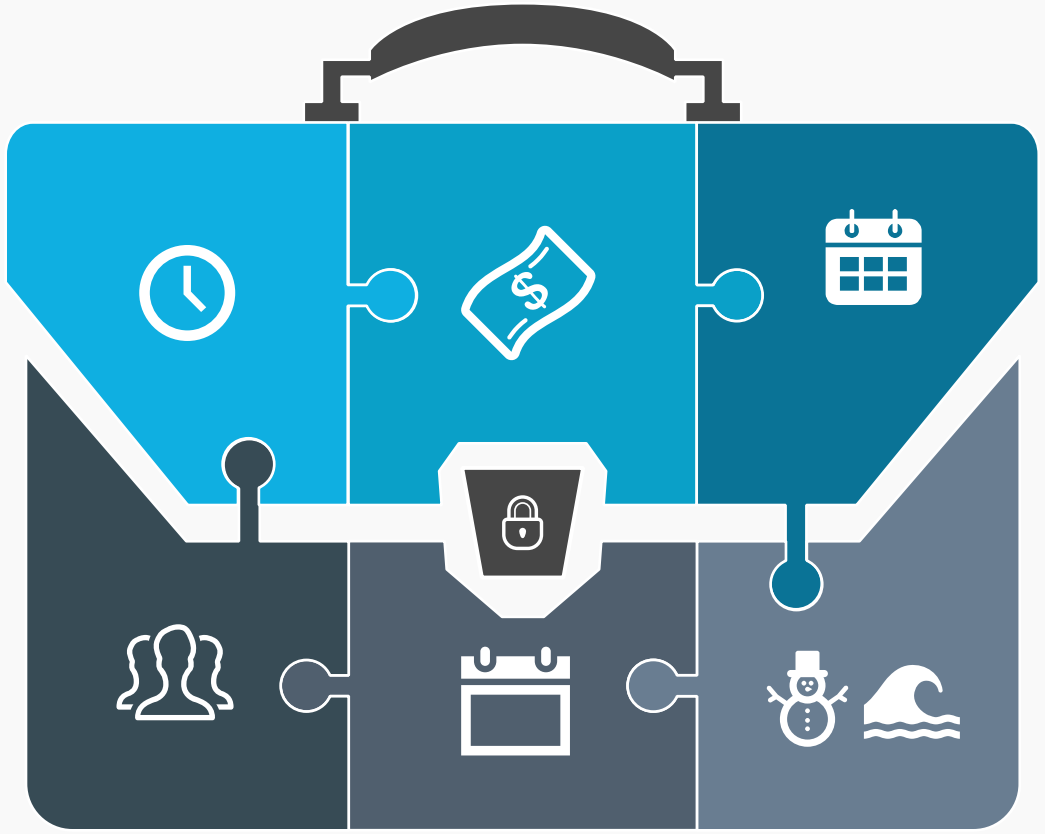
From the previous model selection by GridSearchCV, we see that a larger random forest performs better. So we set the parameter n_estimators to 500.

	RMSE	R^2 score
Decision Tree Regressor	9159.17	0.1330
Random Forest Regressor	9798.6	0.0077

Every classification model are validated using a hold-out test set of 20% the total data

Conclusions

Recomendations Before Launching a Campaign



1. Longer duration between creation and launching
2. Low Project Goals
3. Shorter Campaigns
4. Study the market for competitors
5. Launch your Campaign on Tuesday
6. Avoid launching your campaign during Summer or Christmas.

Lessons Learned

1. How to deal with a large Dataset. We learn how to do an scrapping from a webside, and how to merge several .cvs files into one large data frame using the common attributes.

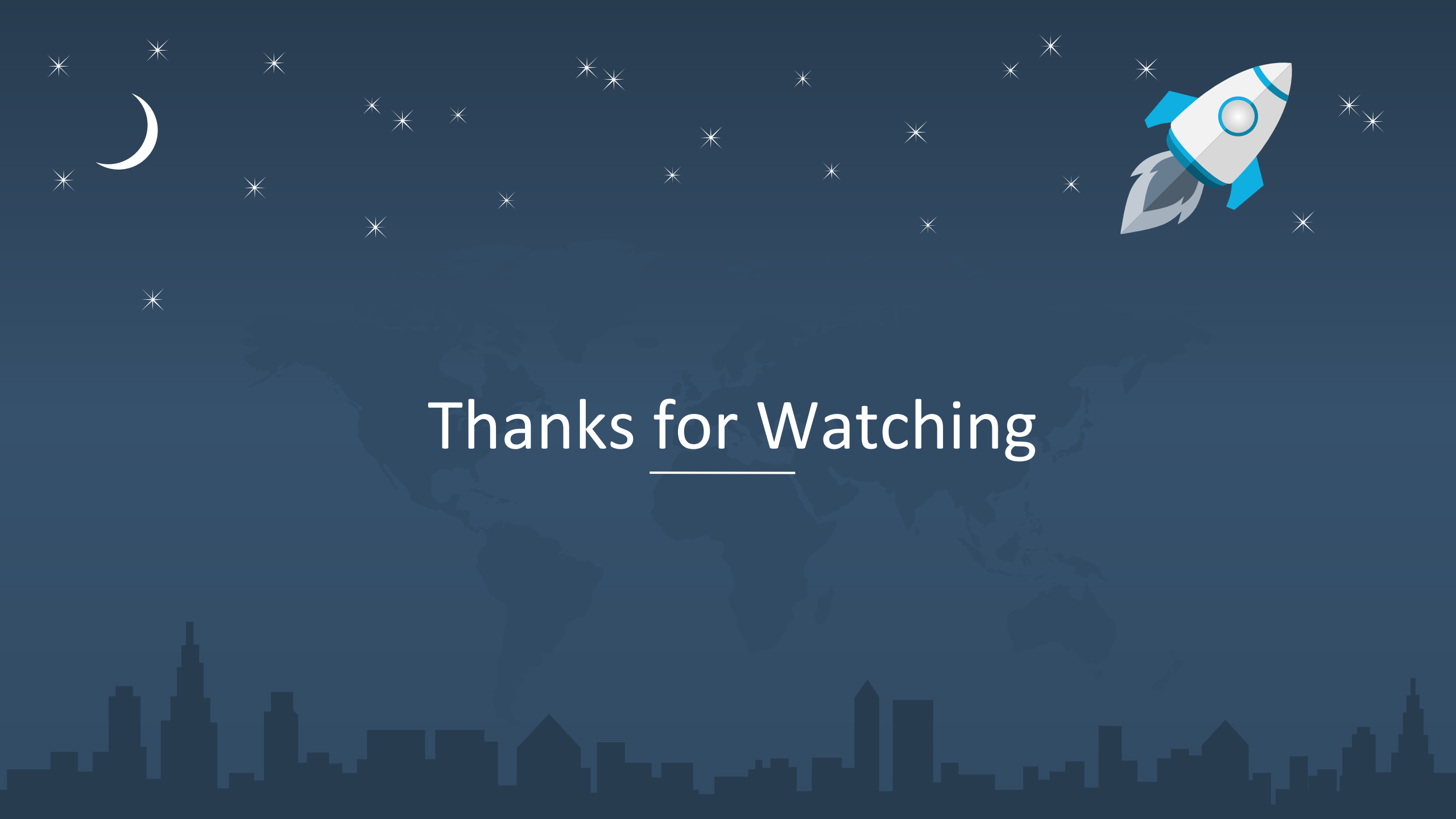
2. In this dataset, we have 250k observations and more than 300 columns. With this large dataset, larger, minimally-pruned decision trees perform very well. In other words, the chance of overfitting the training data is less likely if we have more data.

3. Random Forest is the best performing classification techniques in this project. The more trees we have the better.

4. With the regression problem, decision tree is the better method to use. However, the R^2 of the models is not high.

Next Steps

1. To Include the secondary dataset for improving our model results.
.
2. To finish developing the model deployment and an application Demo.
3. To implement other Machine Learning Models.


Thanks for Watching