SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE
Kolegium Analiz Ekonomicznych

Studia Podyplomowe *„Akademia Analityka z SAS, R & Python”*
Edycja I

Kickstarter – analiza danych i wizualizacja

Imię nazwisko: Bartosz Wiśniowski
Nr albumu: 98901
Praca napisana pod kierunkiem:
Dr inż. Adam Karwan

Warszawa 18.01.2018

Spis treści

1. Struktura pracy

1.1 Cel pracy

Celem napisanej pracy była analiza oraz graficzna wizualizacja danych pochodzących ze zbioru dotyczącego internetowej platformy crowdfundingowej - *Kickstarter*. Dodatkowym założeniem było przetestowanie wybranych metod uczenia maszynowego pod kątem dokładności predykcji sukcesu kampanii prowadzonych na tym serwisie. Numeryczna część pracy w całości została napisana w języku R - z wykorzystaniem środowiska RStudio.

1.2 Zakres pracy

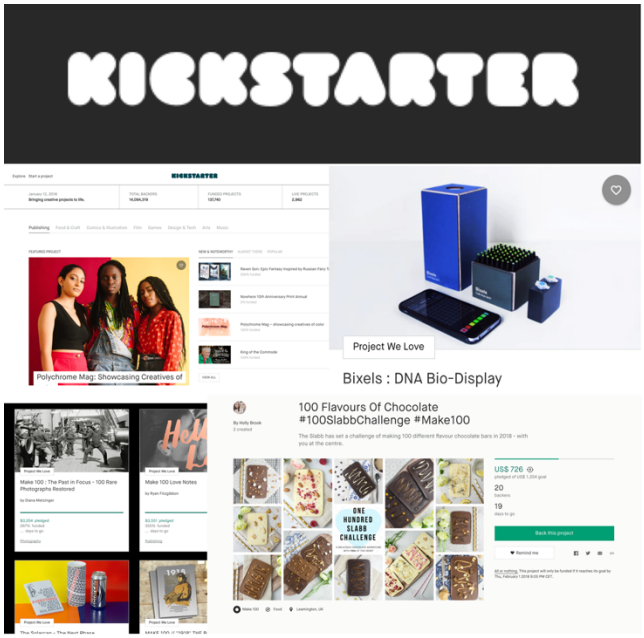
Pierwsza część pracy została poświęcona krótkiemu wprowadzaniu czytelnika do tematyki związanej z serwisem Kickstarter oraz wyjaśnieniem zasady działania crowdfundingu (rozdział 1-2). Następnie skupiono się na szczegółach dotyczących badanego zbioru danych, a także zaprezentowano analizę i wizualizację wybranych, ciekawych z punktu widzenia głównego tematu pracy statystyk (rozdział 3).

W drugiej części pracy (rozdział 4) zaprezentowano zastosowanie czterech podstawowych metod uczenia maszynowego: regresji logistycznej, metody K- najbliższych sąsiadów, metody Maszyny Wektorów Wspierających oraz metody „lasów losowych". Algorytmy zostały przetestowane pod kątem możliwości przewidzenia rezultatu kampanii (sukces lub porażka) na podstawie danych z wcześniej zakończonych przedsięwzięć.

Ostatnia część projektu (rozdziały 5-8) zawiera podsumowanie pracy oraz dokumentację w postaci bibliografii, spisu utworzonych zawartych materiałów oraz kodu napisanego w języku R.

2. Wstęp

Kickstarter – jest amerykańską platformą internetową, umożliwiającą ludziom z całego świata, finasowanie różnego rodzaju kreatywnych przedsięwzięć - od produkcji filmów począwszy, poprzez wydawnictwa książek i gier komputerowych, aż po budowę zaawansowanych technologicznie wynalazków. Każdy z zarejestrowanych na portalu użytkowników ma możliwość stworzenia dla swojego pomysłu kampanii promocyjnej, mającej na celu zachęcenie pozostałych członków społeczności do udzielenia mu wsparcia finansowego, w określonej przez niego wcześniej wielkości. Zasada działania serwisu, opierająca się na zbieraniu małych kwot od dużej grupy ludzi tzw. crowdfunding, pozwala praktycznie każdemu na sfinansowanie nawet najdroższych projektów.



Zdjęcie 1. Kickstarter – przykładowe kampanie

Według danych udostępnionych przez sam serwis, dotychczas przeprowadzonych zostało 139 212 kampanii z których 35.94% zakończyło się sukcesem i w ramach których zebrano około 3.11 miliarda dolarów.

W dalszej części pracy postaram się odpowiedzieć na dwa ciekawe pytania, które nasuwają się po przytoczeniu powyższych statystyk tj. :

- które tak naprawdę parametry mają wpływ na końcowe powodzenie kampanii?
- czy możliwa jest predykcja ewentualnego sukcesu na podstawie danych historycznych?

3. Analiza i wizualizacja danych

3.1 Zbiór danych

Podczas realizacji projektu wykorzystałem zbiór danych pobrany z repozytorium o nazwie *The-Dynamics-of-Rewardbased-Crowdfunding* wykonany przez Roy Klasse w ramach jego pracy licencjackiej i udostępniony przez niego w celach naukowych. Wspomniany zbiór posiada 3652 obserwacje, odpowiadające rzeczywistym kampaniom przeprowadzonym w przeszłości w serwisie Kickstarter, opisane przy pomocy 56 różnorodnych atrybutów.

Wśród wspomnianych cech znajdziemy zarówno te bardzo podstawowe jak np. nazwa projektu promowanego w serwisie, nazwisko twórcy kampanii, zakładany cel finansowy, kategoria przedsięwzięcia, liczba osób która je wsparła czy też informacja o lokalizacji, ale także takie, które pozwalają opisać każdą obserwację na bardziej szczegółowym poziomie jak np. ilość obrazów/zdjęć umieszczona na stronie promocyjnej projektu, wcześniejsze doświadczenia twórcy z serwisem Kickstarter czy też liczba znajomych twórcy na portalu Facebook. W tym miejscu warto dodać, że w celu skutecznej eksploracji danych nie ma konieczności wykorzystywania aż wszystkich atrybutów znajdujących się w zbiorze gdyż niektóre z nich jak np. adres strony www nie wnoszą do rozważań wielu wartościowych informacji. W związku z tym w dalszej części pracy, do wszelkiego rodzaju analiz (w szczególności do tworzenia modeli uczenia maszynowego) wykorzystane zostały wcześniej przygotowane podzbiory danych dotyczących konkretnego zagadnienia.

W celach poglądowych, w tabeli poniżej, zamieszczam jedne z ważniejszych, moim zdaniem, atrybutów ujęte w bardziej ogólne kategorie zbiorcze.

Kickstarter	Finanse	Geografia	Społeczność internetowa	Twórca kampanii
• Kategoria • Liczba wspierających • Ilość zdjęć na profilu • Długość kampanii • Sukces	• Cel finansowy ($) • Wsparcie finansowe ($) • Waluta	• Kontynent • Państwo • Miasto	• Znajomi na Facebook • Ilość „lajków” na portalu Facebook • Liczba komentarzy	• Nazwisko • Doświadczenie na portalu w latach • Ilość wspartych projektów • Liczba przeprowadzonych kampanii

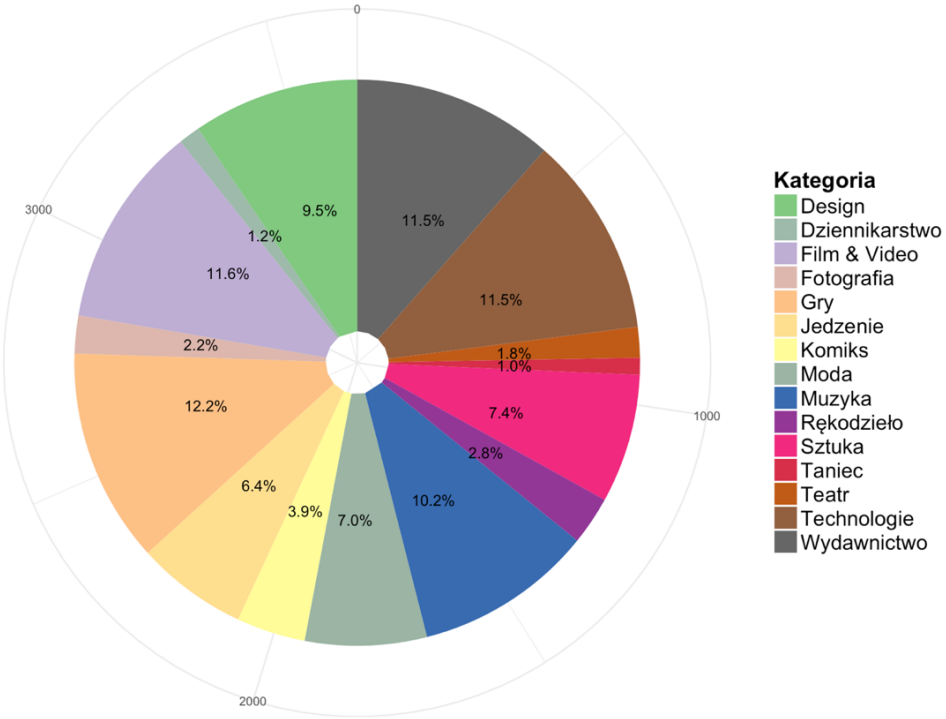
● typ factor ● typ liczbowy

Tabela 1. Przykładowe cechy (atrybuty)

3.2 Wstępna analiza danych

W następującym podrozdziale przeprowadzona została wstępna analiza danych oraz prezentacja statystyk dotyczących wybranych ze zbioru, interesujących atrybutów. Już na etapie początkowej eksploracji danych wyklarowało kilka charakterystycznych atrybutów wydających się mieć największy wpływ na końcowe powodzenie kampanii, należały do nich m.in. kategoria do której należy przedsięwzięcie, czynniki finansowe, położenie geograficzne oraz bezpośrednie aspekty związane z Kickstarterem i twórcą kampanii.

Jako pierwsze zbadane zostały przeze mnie zależności dotyczące występujących kategorii oraz finansów.

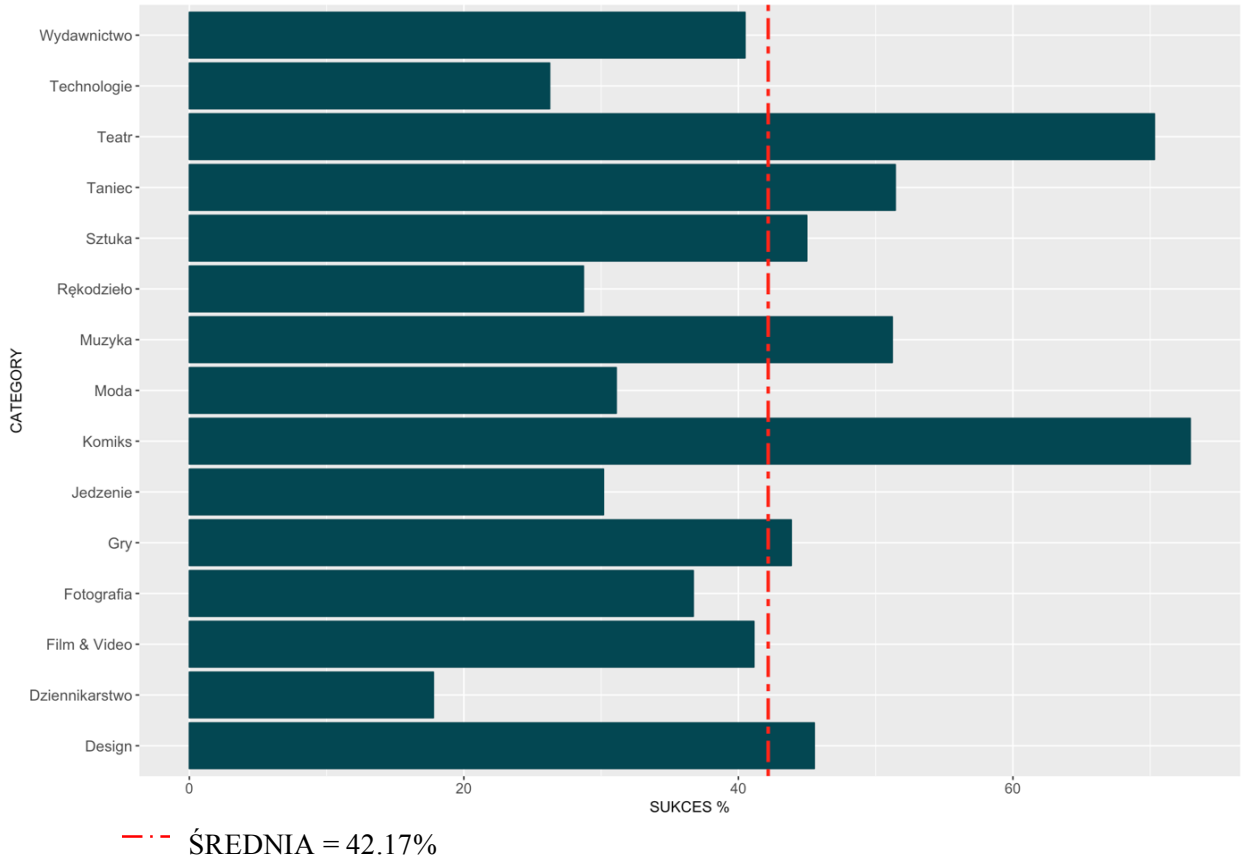


Wykres 1. Procentowy rozkład kampanii wg kategorii

Jak wynika z zaprezentowanego diagramu kołowego najwięcej utworzonych projektów należało do kategorii *Gry* - 12,2% (447 przedsięwzięcia), *Film&Video* - 11,6% (423) oraz *Wydawnictwo* - 11,5% (420) co wydaje oczywiste jeśli wziąć pod uwagę, że reprezentują one przystępną cenowo rozrywką dla mas i kierowane do potencjalnie największej grupy odbiorców. Zaraz za podium znalazła się kategoria *Technologie* 11,5% (419) kampanii, która swoją popularność zawdzięcza pewnie przede wszystkim Dolinie Krzemowej oraz modzie na startupy high - tech.

Na przeciwnym biegunie rankingu popularności znajdują się za to Taniec 1% (35), Dziennikarstwo 1,2% (45) oraz Teatr 1,8% (64). Wynika to prawdopodobnie z faktu, że zarówno przedstawienia teatralne, taneczne jak i dziennikarskie blogi i podcasty, dedykowane są dla wąskiej grupy koneserów.

Skoro już wiadomo jakiego rodzaju przedsięwzięć powstaje najwięcej na całym portalu, w kolejnym kroku należy sprawdzić jak wygląda procentowy sukces kampanii w ramach każdej kategorii.

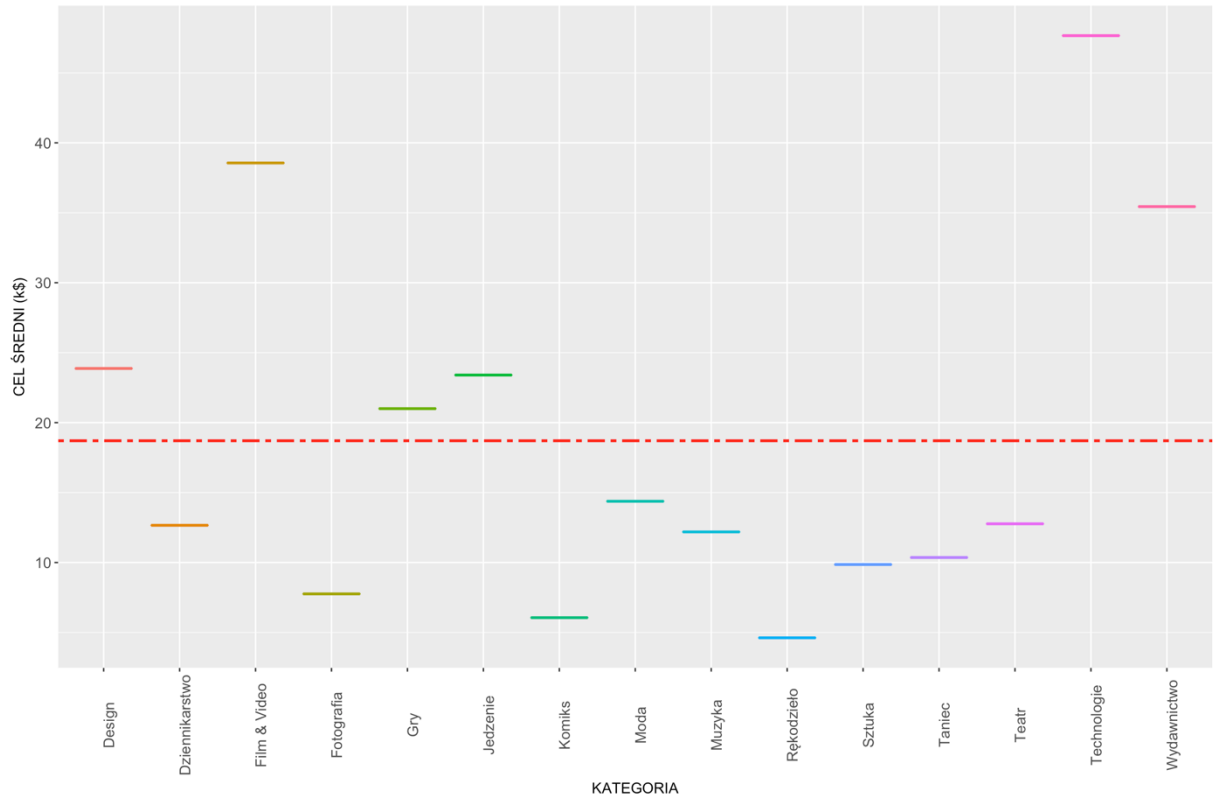


Wykres 2. Procentowy sukces kampanii według danej kategorii

Jak widać na zaprezentowanym powyżej wykresie słupkowym, średni wskaźnik procentowy sukcesu dla dla całego zbioru danych wyniósł 42.17% co oznacza, że finalnie udało się sfinansować mniej niż połowę wszystkich projektów. Najczęściej powodzeniem kończyły się przedsięwzięcia związane z wydaniem komiksu - 72.9% oraz teatrem 70.3% - jednak należy zaznaczyć, że są to kategorie w ramach których przeprowadzono stosunkowo małą liczbę kampanii. Wśród tych dziedzin, które cieszyły się dużą popularnością wśród twórców, najwięcej udanych przedsięwzięć dotyczyło Gier oraz Muzyki i Designu. Na tym etapie rozważań, można

założyć, że to właśnie przynależność do grupy *Gry* jest jedną z cech decydujących o końcowym sukcesie w serwisie Kickstarter.

W kolejnym kroku przeprowadzanej przeze mnie wstępnej analizy danych sprawdziłem jak wyglądał średni zakładany cel finansowy (czyli jak wysoko twórcy wyceniali swoje kampanie) w poszczególnych kategoriach.



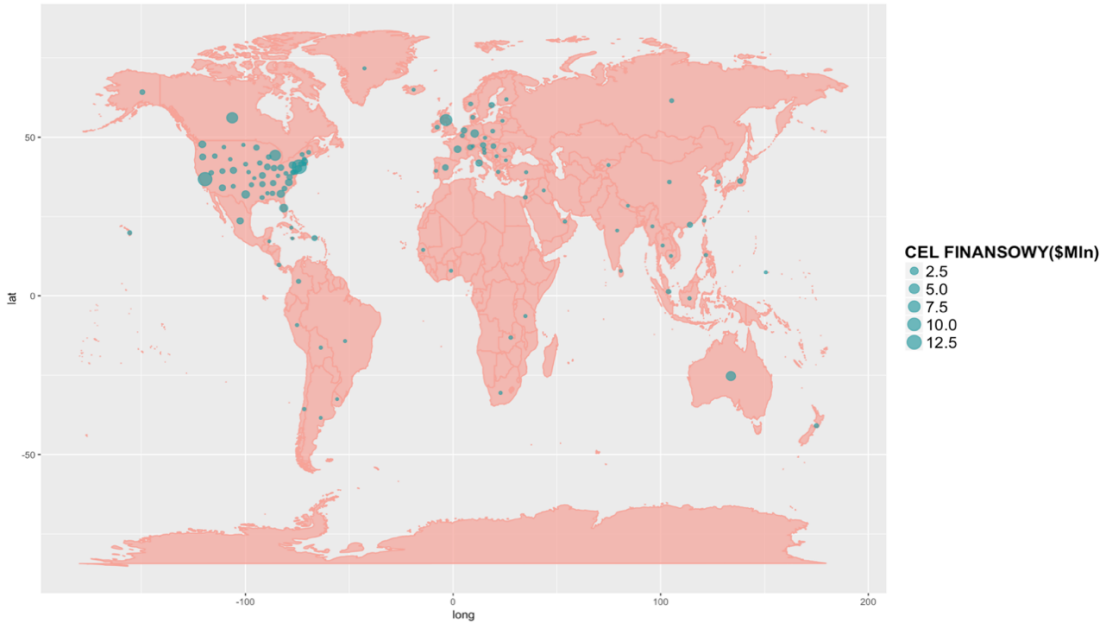
Wykres 3. Średni zakładany cel finansowy dla danej kategorii

Jak wynika z powyższego wykresu, najdroższe kampanie to przede wszystkim te związane z branżą Technologiczną, co wydaje się sensowne gdy weźmiemy pod uwagę jak duże są nakłady finansowe związane z badaniami i rozwojem czy testami produktów. Najmniejsza średnia kwota wymagana do realizacji projektu dotyczyła natomiast kategorii Rękodzieło oraz Komiks (w tym przypadku można zauważyć korelację z procentowym wskaźnikiem sukcesu). Szczegółowe wyniki dla trzech najwyższych i 3 najniższych wyników zamieszczam w tabeli znajdującej się na kolejnej stronie.

Cel finansowy (k$)	
Technologie	**47.66**
Film&Video	38.56
Wydawnictwo	35.44
Fotografia	7.76
Komiks	6.06
Rękodzieło	4.62

Tabela 2. Kategorie o najwyższym oraz najniższym zakładanym celu finansowym

Jako, że Kickstarter jest portalem o zasięgu światowym, warto przyglądnąć się temu jak cechy takie jak wartość kampanii, liczba osób wspierających kampanię oraz całkowite wsparcie finansowe zmieniają się wraz z szerokością geograficzną.

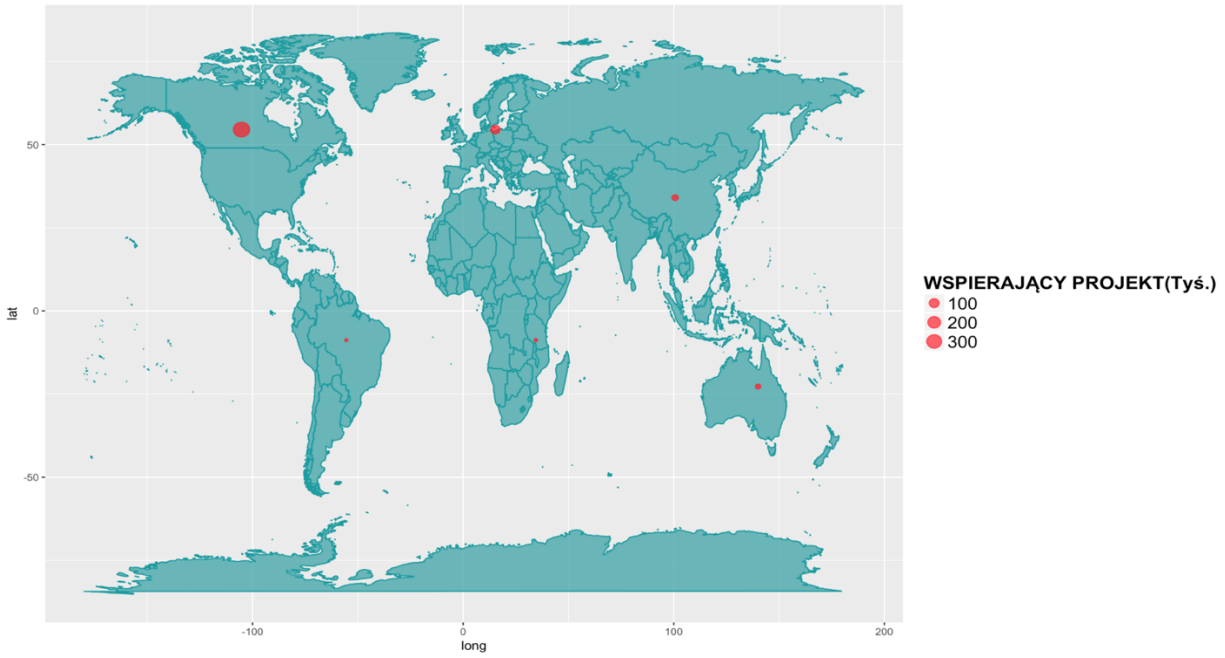


Wykres 4. Geograficzny rozkład wysokości założonego celu finansowego

Na powyższej mapie, można zauważyć, że najdroższe kampanie (czyli te zakładające najwyższy cel finansowy) powstawały na obszarach zachodniego (Kalifornia) oraz wschodniego (Nowy Jork) wybrzeża Stanów Zjednoczonych, Kanady, Wielkiej Brytanii a także Australii. Każde ze wspomnianych miejsc zaliczane jest do bogatych, o wysokim poziomie rozwoju gospodarczego oraz technologicznego. Na duże skupisko tak kosztownych projektów, w tych regionach wpływać

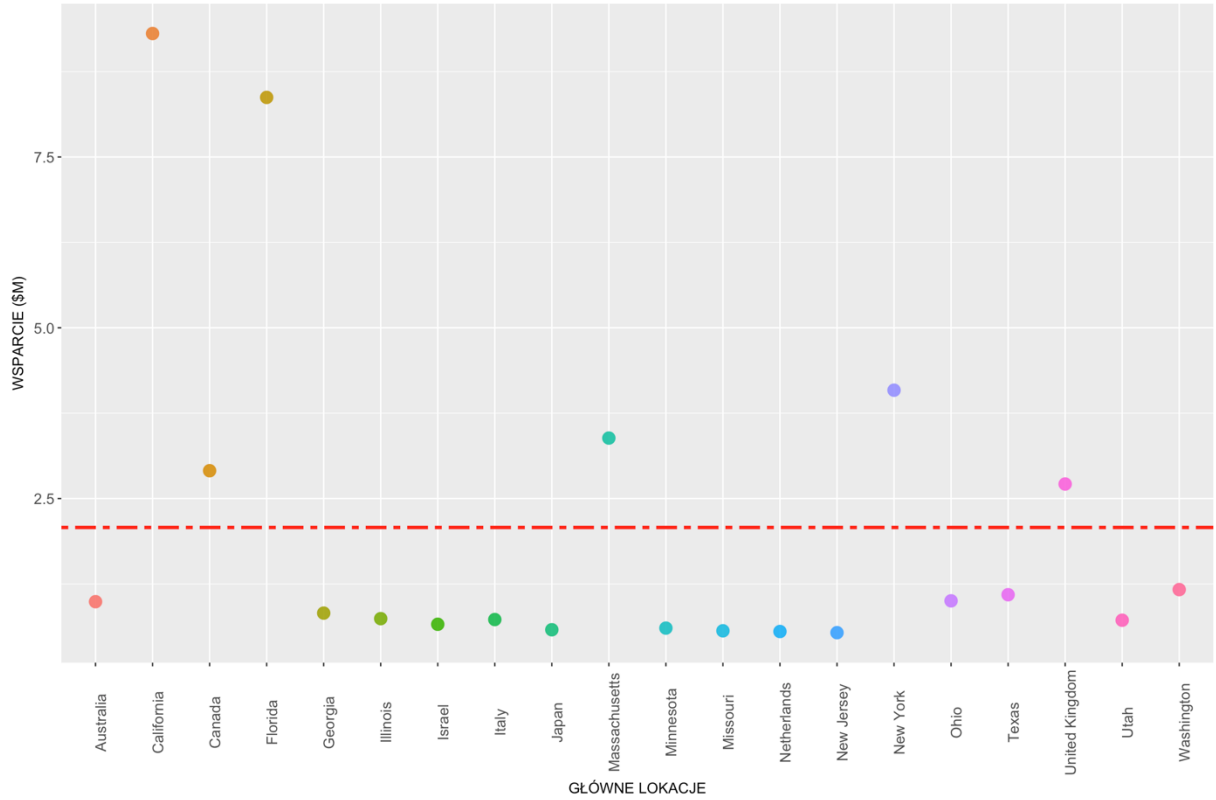
może więc ich przynależność do kategorii Technologie oraz Film (dolina Krzemowa, Hollywood) których to realizacja wymaga największych funduszy.

W podobnym tonie rozkłada się liczba osób inwestujących w projekt. Jak można zauważyć na mapie znajdującej się poniżej, najwięcej użytkowników Kickstartera którzy biorą czynny udział w finansowaniu kampanii zamieszkuje Amerykę Północną oraz Europę a najmniej Amerykę Południową oraz Afrykę. Wpływ na to ma zapewne zarówno status materialny mieszkańców wspomnianych kontynentów a także znajomość języka angielskiego w którym to prowadzony jest portal.



Wykres 5. Geograficzny rozkład wysokości osób wspierających

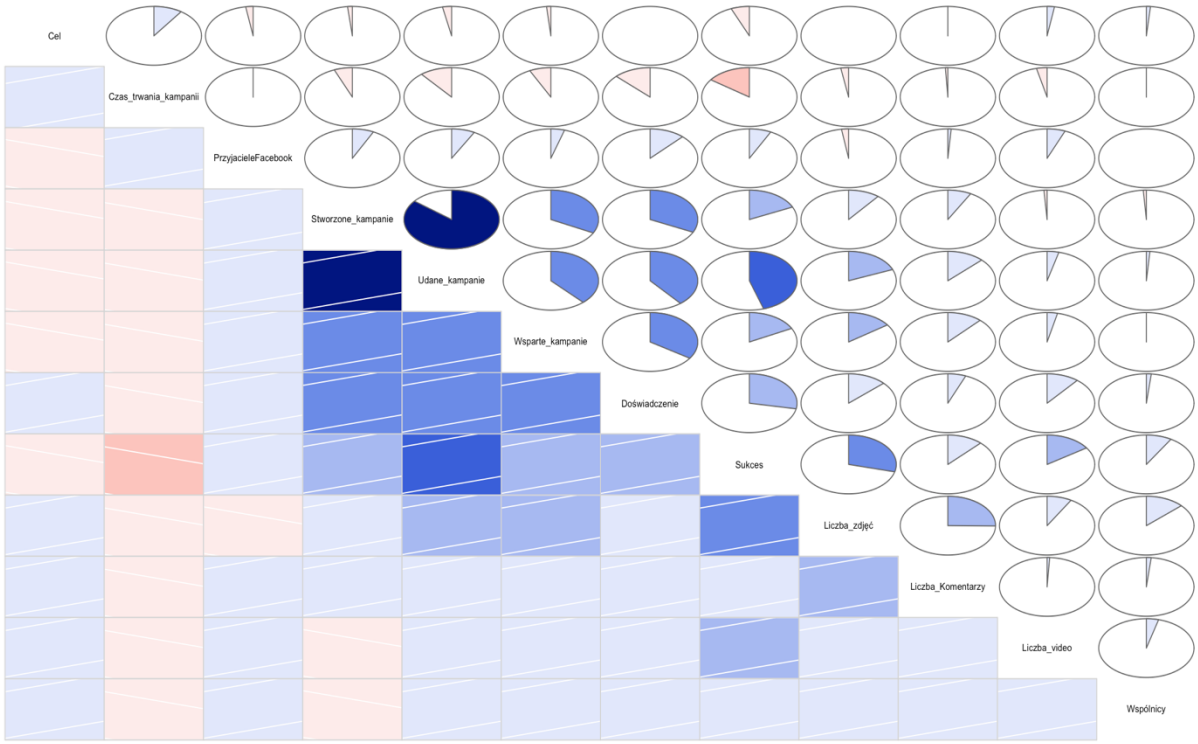
W kolejnym kroku analizy sprawdziłem jak wygląda rzeczywisty rozkład wsparcia finansowego ze względu na szerokość geograficzną a także czy występuje jego korelacja z wcześniej zaprezentowanym rozkładem projektów według celu finansowego.



Wykres 6. Wysokość udzielonego wsparcia finansowego (w mln. dolarów) dla kampanii vs lokalizacja

Jak zostało zaprezentowane na wykresie nr 6, największą ilością pieniędzy zostały wsparte kampanie utworzone w Stanach Zjednoczonych, Kanadzie oraz Wielkiej Brytanii. Rezultat ten pokrywa się z geograficznym rozmieszczeniem kampanii o największej wartości. Wynikać to może zarówno z dużej popularności droższych przedsięwzięć takich jak te z kategorii technologie, jak i po prostu z większej liczby tworzonych projektów w tych regionach.

Ostatnią wyróżnioną grupę atrybutów, znajdujących się w analizowanym przeze mnie zbiorze danych, stanowiły czynniki bezpośrednio związane prowadzeniem kampanii na portalu oraz jej twórcą. Jako, że każda zbiorka funduszy odbywa się wirtualnie, poprzez serwis Kickstarter, można założyć, że odpowiednia kampania promocyjna, doświadczenie oraz poprzednia aktywność twórcy, maja istotny wpływ na ostateczny rezultat. Aby to sprawdzić, w środowisku RStudio, wyznaczyłem macierz korelacji pokazującą bezpośrednie zależności między najważniejszymi atrybutami oraz końcowym sukcesem. Rezultaty zostały przedstawione na korelogramie znajdującym się na kolejnej stronie pracy.



Wykres 7. Korelacja cech dotyczących portalu Kickstarter oraz Internetu

Jak można zauważyć, kluczowy dla nas z punktu widzenia dalszej części pracy, atrybut *Sukces,* największy, dodatni współczynnik korelacji posiada w przypadku atrybutów *Udane_kampanie* oraz *Liczba_zdjęć.* Oznacza to, że większa liczba przeprowadzonych w przeszłości przedsięwzięć zakończonych sukcesem oraz duża liczba zdjęć na profilu promocyjnym kampanii, zwiększają szansę jej twórcy na zgromadzenie funduszy na swój projekt. Na drugim biegunie a więc, atrybutów mających negatywny wpływ na prawdopodobieństwo pomyślnego zakończenia kampanii znajdują się *Czas_trwania_kampanii* oraz *Cel*. Można to zinterpretować tak, że projekty zakładające wyższe cele finansowe oraz te, których kampanie promocyjne trwają dłużej, rzadziej kończą się sukcesem. Należy jednak zaznaczyć, że wszystkie wspomniane przypadki posiadają współczynnik korelacji mniejszy niż 0.5, co według powszechnie uznawanych norm oznacza korelację słabą.

Analiza dokonana w powyższym podrozdziale pozwoliła na bardziej szczegółowe zapoznanie się z omawianym zbiorem danych oraz wstępne określenie głównych parametrów mających wpływ na to jakim rezultatem zakończy się kampania. Aspekt ten zostanie głębiej poruszony w rozdziale następnym, poświęconym wykorzystaniu metod uczenia maszynowego w celach predykcyjnych.

4. Predykcja powodzenia kampanii – analiza skuteczności metod
4.1 Wstęp do modelowania

Jak wspomniałem w rozdziale nr 2, główne pytanie postawione dotyczyło tego czy (oraz z jaką dokładnością) można przewidzieć powodzenie kampanii na podstawie danych archiwalnych, oraz które z atrybutów znajdujących się w zbiorze danych będą najlepszymi predyktorami. W tym celu postanowiłem przetestować cztery popularne metody uczenia maszynowego:

- Regresję logistyczną
- Metodę K – najbliższych sąsiadów (K-Nearest Neighbours)
- Metodę lasów losowych (Random Forrest)
- Metodę Maszyny Wektorów Nośnych (Support Vecotr Machine – SVM)

Pierwszą rzeczą, którą wykonałem w celu optymalizacji całego procesu, było, bazujące na literaturze oraz analizie przeprowadzonej w poprzednim rozdziale, zawężenie grupy cech, które posłużą do budowy modeli predykcyjnych. Wybrane atrybuty znajdują się w tabeli poniżej.

Nazwa	Typ
Cel finansowy	num
Liczba wspartych kampanii	int
Liczba stworzonych kampanii	int
Liczba komentarzy na profilu	num
Doświadczenie na Kickstarter	num
Liczba Kampanii zakończonych sukcesem	int
Liczba współpracowników	int
Czas trwania kampanii	num
Liczba znajomych na Facebook	int
Liczba obrazów zawartych w kampanii	num
Liczba filmów video zawartych w kampanii	num
Sukces	**Factor**

Tabela 3. Atrybuty wykorzystane przy budowie modeli predykcyjnych

Kolejnym krokiem na tym etapie był podziału zbioru na dwa podzbiory:

- treningowy - który służył do nauczenia wybranych algorytmów
- testowy - na którym mogłem sprawdzić dokładność moich metod

Zdecydowałem się podzielić zbiór wejściowy w stosunku 70%-30% (trening- test).

4.2 Metoda regresji logistycznej

Pierwszą z zaimplementowanych przeze mnie metod była regresja logistyczna. Jest to algorytm wykorzystywany w statystyce, wtedy, gdy zmienna zależna jest zmienną dychotomiczną czyli przyjmującą tylko dwie wartości (w moim przypadku był to atrybut *Sukces* , który przyjmuje 0 lub 1). Po stworzeniu modelu klasyfikacji w środowisku RStudio a następnie przetestowaniu go na zbiorze testowym otrzymałem macierz konfuzji która pokazuje, że poprawnie wytypowano 628 kampanii zakończonych sukcesem oraz 445 kampanii zakończonych porażką.

	TRUE	FALSE
Sukces	**628**	**17**
Porażka	**6**	**445**

Tabela 4. Macierz konfuzji dla algorytmu regresji logistycznej

Na podstawie wspomnianej macierzy można określić podstawową miarę jakości klasyfikacji czyli **Dokładność (Accuracy).** W moim przypadku jest to:

Accuracy = (628+445)/(628+445+6+17)= **0.979**

Dzięki zastosowaniu, będącej częścią języka R, funkcji summary(), wyznaczyłem także predyktory o największym wpływie na zbudowany model. Jak pokazuje poniższa tabela, „waga" atrybutów pokrywa się z wynikami zaprezentowanymi na wcześniejszym korelogramie.

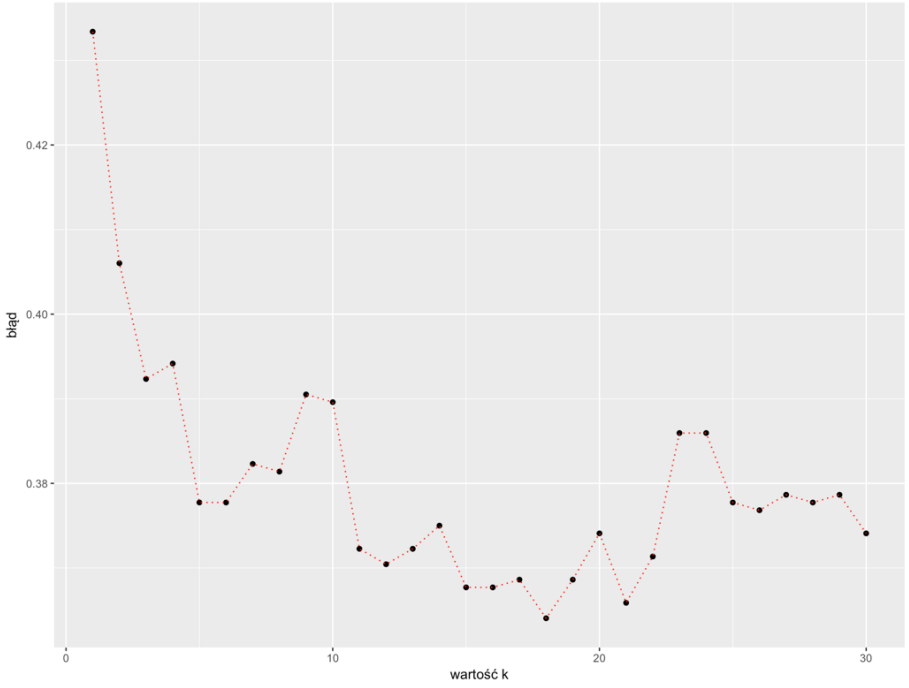
Atrybut
Cel finansowy
Liczba stworzonych kampanii
Liczba wygranych kampanii
Liczba komentarzy na profilu

Tabela 5. Predyktory o najwyższej wadze wyznaczone dla modelu regresji logistycznej

4.3 Metoda K – najbliższych sąsiadów

Kolejnym z przetestowanych algorytmów była metoda „najbliższych k sąsiadów" (*K-Nearest Neighbours* lub *KNN*). Algorytm ten przyporządkowuje obiekt do danej klasy (opisanej szeregiem cech) poprzez wyszukanie w jego otoczeniu k najbliższych obiektów dla których klasa jest znana, i wybranie tej najbardziej licznej.

Po zaimplementowaniu modelu dla algorytmu KNN w RStudio (kod napisany w języku R został załączony do pracy), w celu znalezienia optymalnej liczby sąsiadów k, wygenerowane zostały wartości błędu klasyfikacji dla 30 różnych parametrów k (w przedziale od 1 do 30).



Wykres 8. Parametr k vs błąd klasyfikacji

Na wykresie nr 8 można zauważyć, że najmniejszy błąd klasyfikacji **e=0.364** otrzymaliśmy dla parametru **k=18** (czyli otoczenia obejmującego 18 sąsiadów). Daje to dokładność predykcji rzędu **64,6%** czyli znacznie niższą niż w przypadku poprzednio pokazanej regresji logistycznej.

4.4 Algorytm „lasów losowych"

Trzecią z metod wykorzystanych w celu zbudowania modelu predykcyjnego był algorytm „lasów losowych" (*Random Forrest*). Zasada jego działania opiera się na utworzeniu wielu drzew decyzyjnych dla każdego z których, podzbiór analizowanych cech w węźle, dobierany jest losowo. Bazując na literaturze, algorytm ten powinien wykazać się najwyższą dokładnością

z pośród wszystkich testowanych. Po zaimplementowaniu algorytmu w środowisku RStudio (kod znajduje się w załączniku do pracy) otrzymałem następujące rezultaty:

	TRUE	FALSE
Sukces	**630**	**3**
Porażka	**15**	**448**

Tabela 6. Macierz konfuzji dla algorytmu Random Forrest

Dokładność klasyfikacji dla zastosowanego algorytmu oraz zadanego zbioru testowego wynosi:

Accuracy = (630+448)/(630+448+3+15)= **0.974**

4.5 Metoda Maszyny Wektorów Wspierających

Ostatnim z zastosowanych algorytmów jest metoda „Maszyny Wektorów Wspierających" (*Support Vector Machine* lub *SVM*). Jest to skuteczny klasyfikator, którego zasada działania, w uproszczeniu, opiera się na wyznaczeniu tzw. hiperpłaszczyzn rozdzielających (z maksymalnym marginesem) obiekty należące do różnych klas. Poniżej zamieszczam macierz konfuzji, którą otrzymałem po zaimplementowaniu algorytmu *SVM* w środowisku RStudio (kod w języku R znajduje się w załączniku do pracy).

	TRUE	FALSE
Sukces	**612**	**5**
Porażka	**33**	**446**

Tabela 7. Macierz konfuzji dla algorytmu SVM

Dokładność klasyfikacji dla zastosowanego algorytmu wynosi:

Accuracy = (612+446)/(612+446+33+5)= **0.965**

Podobnie jak w przypadku w przypadku regresji logistycznej oraz metody lasów losowych jest to wysoka, ponad 90% dokładność.

5. Podsumowanie

W zrealizowanej pracy zawarte zostały wyniki analizy statystycznej i graficzne wizualizacje danych pochodzących ze zbioru dotyczącego archiwalnych kampanii przeprowadzonych na portalu Kickstarter. Na podstawie przeanalizowanych danych wnioskować można, że:

- Do najbardziej popularnych kategorii na portalu Kickstarter należą: Gry, Film & Video, oraz Wydawnictwo a do najmniej Teatr i Taniec
- Najdroższe kampanie powstają w Ameryce Północnej oraz Wielkiej Brytanii – na co wpływ ma zapewne wysoki poziom gospodarczy i technologicznych tych regionów a także status materialny zamieszkujących tam ludzi
- Sukces kampanii zależy także w dużym stopniu od wcześniej nabytego doświadczenia i umiejętności twórcy kampanii (duże znaczenie ma tu liczba wcześniej stworzonych projektów a przede wszystkim liczba projektów zakończonych sukcesem), od zakładanego celu finansowego a także od liczby komentarzy na profilu
- Swoje szanse na sukces można zwiększyć tworząc projekt w kategorii Gry

Ponad to na bazie rezultatów uzyskanych w rozdziale 4, udowodniono, że możliwa jest predykcja sukcesu kampanii poprzez wykorzystanie danych archiwalnych oraz metod metod uczenia maszynowego. Przeprowadzone testy wykazały ze najwyższą dokładność predykcji otrzymaliśmy dla algorytmu regresji logistycznej a najmniejszą dla metody k-najbliższych sąsiadów. Pełne zestawienie dla wszystkich czterech algorytmów zamieszone zostało w tabeli poniżej.

Algorytm	Dokładność
Regresja logistyczna	**97,9%**
Lasy losowe	**97,4%**
SVM	**96,5%**
KNN	**64,6%**

Tabela 8. Ranking przetestowanych algorytmów

6. Bibliografia

1. https://github.com/RoyKlaasseBos/The-Dynamics-of-Rewardbased-Crowdfunding
2. https://towardsdatascience.com/predicting-the-success-of-kickstarter-campaigns-3f4a976419b9
3. http://wojtkiewicz.eu/wp-content/uploads/2014/10/Algorytm-Random-Forest-.pdf
4. https://pl.wikipedia.org/wiki/K_najbli%C5%BCszych_s%C4%85siad%C3%B3w
5. http://www.staff.amu.edu.pl/~drizzt/images/DSSU/W8.pdf
6. https://www.kickstarter.com/help/stats
7. https://www.nemoursresearch.org/open/StatClass/January2011/Class8.pdf
8. http://web.mit.edu/zoya/www/SVM.pdf
9. http://mathspace.pl/matematyka/confusion-matrix-macierz-bledu-tablica-pomylek-czyli-ocena-jakosci-klasyfikacji-czesc-1/

7. Spis zawartości

7.1 Spis wykresów

7.2 Spis tabel

7.3 Spis wykorzystanych zdjęć

8. Kod w języku R

```
#Wczytywanie i sprawdzanie danych
data=csv.read("kick.csv")
colSums(is.na(kick_raw))
summary(data)
str(data)

#Agregacja wg krajów
countries=aggregate(totalPledge ~ countryState, data, sum)
countries$totalPledge=countries$totalPledge/1000000
countries$goal=aggregate(goal ~ countryState, data, sum)[[2]]
countries$goal=countries$goal/1000000

#Piechart
library(RColorBrewer)
p = ggplot(data.kategorie, aes(x=factor(1),y=Suma, fill = factor(Kategoria)) )
p=p + geom_bar( stat = "identity", width=0.8)
p=p+ geom_text(aes(label =data.kategorie$kat_procent1),position=position_stack(vjust=0.6) )
p=p+scale_fill_manual(values = colorRampPalette(brewer.pal(8,"Accent"))(15))
p=p+ coord_polar(theta="y",start=0)
p=p+xlab('') + ylab('') + labs(fill='Kategoria')
p=p+ theme(legend.title = element_text(size = 16,, face = "bold"),legend.text = element_text(size = 15))
print(p)

#Liczba projektów vs Kategorie
gg=ggplot(data2, aes(mainCategory)) +
geom_bar(width= 0.9,color="#034752", fill="#034752") +
 theme_minimal()+ theme(axis.text.x = element_text(size =9,angle=90), axis.text.y = element_text(size
=9),legend.position="none") + labs(x="CATEGORY", y="PROJECTS")

#Liczba wspierających vs Kategorie
gg= ggplot(data, aes(mainCategory))+
geom_bar(aes(weight=data$totalNumberBackers),fill="#034752",color="#034752") +
theme_minimal()+theme(axis.text.x = element_text(size =10,angle=90),legend.position="none") +
labs(x="Main Categories", y="Total number of backers")+
geom_hline(yintercept = mean(kategorie$Cel_średni), color="#ff1919", linetype="twodash", size=1.2)

#Kategoria vs Średni cel finansowy
gg= ggplot(data.kategorie, aes(x=Kategoria, y=Cel_średni)) +
geom_point(aes(color=kategorie$Kategoria),shape=16,alpha=0.9,size=3,stroke =1)+
theme_update()+ labs(x="KATEGORIA",y="CEL ŚREDNI (k$)")+
theme(axis.text.x = element_text(size =11,angle=90),axis.text.y = element_text(size =11),
legend.position="none")+
geom_hline(yintercept = mean(kategorie$Cel_średni), color="#ff1919", linetype="twodash", size=1.2)

#Wsparcie vs Główne lokalizacje
gg= ggplot(topCountries, aes(countryState,totalPledge))+ geom_point(aes(color=
topCountries$totalPledge),shape=16,alpha=0.9,size=3,stroke =1)+ theme_update()+theme(axis.text.x =
element_text(size =11,angle=90), axis.text.y = element_text(size =11), legend.position="none") +
labs(x="Główne lokacje", y="Wsparcie"
```

#CelFinansowy vs Lokalizacja

```
library("ggmap")
library(maptools)
library(maps)
visited <- as.character(topCountries$countryState)
ll.visited <- geocode(visited)
visit.x <- ll.visited$lon
visit.y <- ll.visited$lat
mapWorld <- borders("world", colour="#1d9da1", fill="#1d9da1",alpha=0.9) # create a layer of borders
mp <- ggplot() +   mapWorld
mp <- mp+ geom_point(aes(x=visit.x, y=visit.y, size=topCountries$totalPledge)
,color="#ff8a28",alpha=0.7) + scale_fill_manual(name="Top Locations/Pledges",
 labels=sprintf("%fmln",topCountries$totalPledge/1000000))
```

#Kontynent vs Wspierajacy

```
library("ggmap")
library(maptools)
library(maps)
visited <- as.character(continents$continent)
ll.visited <- geocode(visited)
visit.x <- ll.visited$lon
visit.y <- ll.visited$lat
mapWorld <- borders("world", colour="#1d9da1", fill="#1d9da1",alpha=0.9) # create a layer of borders
mp <- ggplot() +   mapWorld
mp <- mp+ geom_point(aes(x=visit.x, y=visit.y, size=continents$backers) ,color="#ff8a28",alpha=0.7)
mp
```

#Korelogram atrybutów związanych z internetem

```
library(corrgram)
data.internet=select(data,goal,numComments,backed,created,experience,facebookFriends,numCollabor
ators, numSuccessfulCampaigns, mainVideo,numImages,durationCampaign,success)
colnames(data.internet)=c('Cel','Liczba_Komentarzy','Wsparte_kampanie','Stworzone_kampanie','Doświ
adczenie','PrzyjacieleFacebook','Wspólnicy','Udane_kampanie','Liczba_video','Liczba_zdjęć','Czas_trwani
a_kampanii','Sukces')
corrgram(data.internet, order=TRUE, lower.panel=panel.shade, upper.panel=panel.pie,
text.panel=panel.txt)
```

#Przetworzenie danych przed zastosowaniem algorytmów ML

```
library(dplyr)
data.prediction2=select(data,goal,numComments,backed,created,
experience,facebookFriends,numCollaborators, numSuccessfulCampaigns,
mainVideo,numImages,durationCampaign,success)
data.prediction1$success=factor(data.prediction1$success)
```

#Podział na zbioru na treningowy i testowy

```
library(caTools)
set.seed(101)
spl = sample.split(data.prediction2$success, 0.7)
train = subset(data.prediction2, spl == TRUE)
test = subset(data.prediction2, spl == FALSE)
```

```
#Metoda regresji logistycznej
library(caTools)
model = glm(success ~ ., family = binomial(logit), data = train)
summary(model)
#new.step.model <- step(model)
#summary(new.step.model)
test$predicted.success = predict(model, newdata=test, type="response")
table(test$success, test$predicted.success> 0.5)

#SVM
library(e1071)
model <- svm(success ~ .,data=train)
summary(model)
predicted.values <- predict(model,test[1:13])
table(predicted.values,test$success)

#Metoda lasów losowych
library(randomForest)
rf.model <- randomForest(success ~ . , data = train,importance = TRUE)
rf.model$confusion
rf.model$importance
p <- predict(rf.model,test)
table(p,test$success)

#KNN
library(class)
predicted.success <- NULL
error.rate <- NULL
for(i in 1:10)
{set.seed(101)
 predicted.success <- knn(train[1:13], test[1:13], train$success, k=i)
 error.rate[i] <- mean(test$success != predicted.success) }
mean(test$success != predicted.success)
library(ggplot2)
k.values <- 1:30
error.df <- data.frame(error.rate,k.values)
pl <- ggplot(error.df,aes(x=k.values,y=error.rate)) + geom_point()
pl + geom_line(lty="dotted",color='red')
```