Data Science 1 Projekt Report Anhang

Zahra Abdi 6350372 Christopher Grosse Aylin Demir
6350372 6570589 6986267

1 Angepasster Datensatz

Spaltennamen Datensatz Happiness mit Relevanzeinstufung in Bezug zum Projekt:

Spaltenname	Relevant	Ggf. zukünftig relevant	Irrelevant
Gesamte Anzahl	3	9	0
Country	X		
Happiness Rank	X		
Happiness Score	X		
Whisker.high		X	
Whisker.low		X	
Economy GDP per Capita		X	
Family		X	
Health Life Expectancy		X	
Freedom		X	
Generosity		X	
Trust Government Corruption		X	
Dystopia Residual		X	

Tabelle 1: Happiness mit Relevanzeinstufung für das Projekt

Spaltennamen Datensatz Freedom mit Relevanzeinstufung in Bezug zum Projekt:

Spaltenname	Relevant	Ggf. zukünftig relevant	Irrelevant
Gesamte Anzahl	6	104	10
year	X		
ISO_code	X		
countries	X		
region	X		
hf_score	X		
hf_rank	X		
hf_quartile		X	
pf_rol_procedural		X	
pf_rol_civil		X	
pf_rol_criminal		X	
pf_rol		X	
pf_ss_homicide		X	
pf_ss_disappearances_disap		X	
pf_ss_disappearances_violent		X	
pf_ss_disappearances_organized		X	
pf_ss_disappearances_fatalities		X	
pf_ss_disappearances_injuries		X	
pf_ss_disappearances		X	
pf_ss_women_fgm		X	
pf_ss_women_inheritance_widows			X
pf_ss_women_inheritance_daughters			X

pf_ss_women_inheritance		X	
pf_ss_women		X	
pf_ss		X	
pf_movement_domestic		X	
pf_movement_foreign		X	
pf_movement_women		X	
pf_movement		X	
pf_religion_estop_establish			X
pf_religion_estop_operate			X
pf_religion_estop		X	
pf_religion_harassment		X	
pf_religion_restrictions		X	
pf_religion		X	
pf_association_association		X	
pf_association_assembly		X	
pf_association_political_establish			X
pf_association_political_operate			X
pf_association_political		X	
pf_association_prof_establish			X
pf_association_prof_operate			X
pf_association_prof		X	
pf_association_sport_establish			X
pf_association_sport_operate			X
pf_association_sport		X	
pf_association		X	
pf_expression_killed		X	
pf_expression_jailed		X	
pf_expression_influence		X	
pf_expression_control		X	
pf_expression_cable		X	
pf_expression_newspapers		X	
pf_expression_internet		X	
pf_expression		X	
pf_identity_legal		X	
pf_identity_sex_male		X	
pf_identity_sex_female		X	
pf_identity_sex		X	
pf_identity_divorce		X	
pf_identity		X	
pf_score		X	
pf_rank		X	
ef_government_consumption		X	
ef_government_transfers		X	
ef_government_enterprises		X	
ef_government_tax_income		X	
ef_government_tax_payroll		X	
ef_government_tax		X	
ef_government_soa		X	
ef_government		X	
ef_legal_judicial		X	
ef_legal_courts		X	

ef_legal_protection		X	
ef_legal_military		X	
ef_legal_integrity		X	
ef_legal_enforcement		X	
ef_legal_restrictions		X	
ef_legal_police		X	
ef_legal_crime		X	
ef_legal_gender		X	
ef_legal		X	
ef_money_growth		X	
ef_money_sd		X	
ef_money_inflation		X	
ef_money_currency		X	
ef_money		X	
ef_trade_tariffs_revenue		X	
ef_trade_tariffs_mean		X	
ef_trade_tariffs_sd		X	
ef_trade_tariffs		X	
ef_trade_regulatory_nontariff		X	
ef_trade_regulatory_compliance		X	
ef_trade_regulatory		X	
ef_trade_black		X	
ef_trade_movement_foreign		X	
ef_trade_movement_capital		X	
ef_trade_movement_visit		X	
ef_trade_movement		X	
ef_trade		X	
ef_regulation_credit_ownership		X	
ef_regulation_credit_private		X	
ef_regulation_credit_interest		X	
ef_regulation_credit		X	
ef_regulation_labor_minwage		X	
ef_regulation_labor_firing		X	
ef_regulation_labor_bargain		X	
ef_regulation_labor_hours		X	
ef_regulation_labor_dismissal		X	
ef_regulation_labor_conscription		X	
ef_regulation_labor		X	
ef_regulation_business_adm		X	
ef_regulation_business_bureaucracy		X	
ef_regulation_business_start		X	
ef_regulation_business_bribes		X	
ef_regulation_business_licensing		X	
ef_regulation_business_compliance		X	
ef_regulation_business		X	
ef_regulation		X	
ef_score		X	
ef_rank		X	

Tabelle 2: Freedom mit Relevanzeinstufung für das Projekt

Liste der veränderten Ländernamen:

Ländernamen - Freedom Datensatz		Ländernamen - Happiness – Datensatz	
Korrigierter Ländername	Original Datensatzname	Korrigierter Ländername	Original Datensatzname
Central Afr. Rep.	Central African Republic	-	-
Czech Rep.	Czech Republic	-	-
Democratic Republic of Congo	Congo, Dem. R.	Democratic Republic of Congo	Congo (Kinshasa)
-	-	Hong Kong	Hong Kong S.A.R.
Ivory Coast	C?te d'Ivoire	-	-
Yemen	Yemen, Rep.	-	-
Kyrgyzstan	Kyrgyz Republic	-	-
-	-	North Macedonia	Macedonia
Republic of Congo	Congo, Rep. Of	Republic of Congo	Congo (Brazzaville)
Slovakia Republic	Slovak Rep.	Slovakia Republic	Slovakia
South Korea	Korea, South	-	-
-	-	Taiwan	Taiwan Province of China

Tabelle 3: Happiness mit Relevanzeinstufung für das Projekt

Liste der Ländernamen mit fehlendem Happiness-Wert oder Freedom-Wert:

Ländernamen fehlender Freedom-Wert	Ländernamen fehlender Happiness-Wert
Afghanistan	Bahamas
Kosovo	Barbados
Palestinian	Brunei Darussalam
Somalia	Cape Verde
South Sudan	Eswatini
Turkmenistan	Fiji
Uzbekistan	Gambia
	Guinea Bissau
	Guyana
	Laos
	Oman
	Pap. New Guinea
	Seychelles
	Suriname
	Timor Leste

Tabelle 4: Ländernamen mit fehlendem Happiness-Wert oder Freedom-Wert

Liste der Ländernamen, bei denen ein Komma innerhalb des Ländernamens entfernt wurde:

Originaler Datensatzname	Korrigierter Ländername
Gambia, The	Gambia

Tabelle 5: Ländernamen, bei denen ein Komma entfernt wurde

Zusammenführung von den Werten von Nordzypern (North Cyprus) und Zypern (Cyprus) im Happiness Datensatz:

	Bevölkerung	Prozentsatz
Cyprus	1120489	0,7496
North Cyprus	374299	0,2504

Die neuen Werte:

Happiness.Rank	63.9984
Happiness.Score	5.6676
Whisker.high	5.761
Whisker.low	5.576
Economy GDP	1.3537
Family	1.1452
Health	0.8422
Freedom	0.3842
Generosity	0.2702
Trust	0.0707
Dystopia	1.6033

Tabelle 6: Zusammenführung von den Werten von Nordzypern und Zypern

2 SQL-Code

Der SQL-Code für das Mergen von den beiden Datensätzen:

```
SELECT
t1.ISO_code as ISO_Code,
t1.countries as Country_FD, t1.region, t1.hf_score as Freedom_Score,
t2.Country as Country_HN, t2.Happiness_Score as Happiness_Score /*t4.Bundesland as BLG3, t2.BLG4,
t2.BLG5, t2.BLG6, t2.BLG7, t2.BLG8, t2.BLG9*/ FROM 1datascience.freedom2017 t1
LEFT JOIN (SELECT Country, Happiness_Score
FROM 1datascience.happiness) t2 on t1.countries = t2.Country
Where Happiness_Score is not null;
```

3 Cluster Algorithmen

3.1 K-means Algorithmus

Der K-Means Algorithmus ist ein Cluster-Algorithmus, der Objekte beziehungsweise Daten, die eine Ähnlichkeit aufweisen, in Cluster setzt. Für den Algorithmus wird von dem Entwickler ein vordefinierter Wert k angegeben. Dieser gibt die Anzahl von Clusterzentren und somit auch die Anzahl von Clustern an.
Jedes Objekt wird dann dem jeweiligen nächsten Zentrum, der beim ersten Durchlauf zufällig gewählt wird, zugewiesen. Daraufhin werden die Zentren neu berechnet, indem der Mittelwert der Koordinaten von den Datenpunkten im Cluster berechnet wird. Dies wird getan, da die Zentren zu Anfangs nicht optimal auf die Daten verteilt werden, sodass es zum Beispiel passieren könnte, dass alle Punkte ein einziges Zentrum als das nächstgelegene ansehen und die anderen Zentren keine Daten zugeordnet bekommen.
Wurde die Zentren neu gewählt, werden die Daten wieder den neuen Zentren zugeteilt. Dies wird solange durchgeführt, bis die Abbruchbedingung stattfindet. Die Abbruchbedingung sieht vor, dass der Algorithmus beendet ist, wenn sich die Zentren nach zwei Iterationen nicht unterscheiden und somit die finalen Zentren gefunden wurden.
Der k-means ist ein effizienter Algorithmus, da seine Laufzeit linear zur Anzahl der Datenpunkte ist.

3.2 Affinity Propagation Algorithmus

Im Gegensatz zum k-means Algorithmus wird beim Affinity Propagation Algorithmus die Anzahl von Cluster selbst berechnet.
Durch das Senden von Nachrichten zwischen Paaren von Samples, bis eine Konvergenz entsteht, werden Cluster gebildet.
Der Datensatz wird durch einige Exemplare beschrieben, die als die repräsentativsten Daten identifiziert werden. Die Nachrichten, die zwischen den Paaren gesendet werden, repräsentieren die Einigung für ein Objekt als Exemplar eines anderen Objekts, das wegen dem Wert von einem anderem Paar aktualisiert wurde. Diese Aktualisierung wird iterativ durchgeführt, bis zur Konvergenz. Dadurch werden die finalen Exemplare ausgewählt und somit auch die Cluster.
Der Affinity Propagation Algorithmus hat eine Laufzeit von $O(N^2T)$, wobei N die Anzahl von Objekten und T die Anzahl von den Iterationen bis zur Konvergenz sind.