Data Cleaning and Analysis: Generating New Datasets and Visualizing Results

Cleaning

1- Duplicate :

duplicate detection before insertion in the database

duplicates will be removed on joining tables

2- Unifying formats to UTF 8

Generating New Dataset

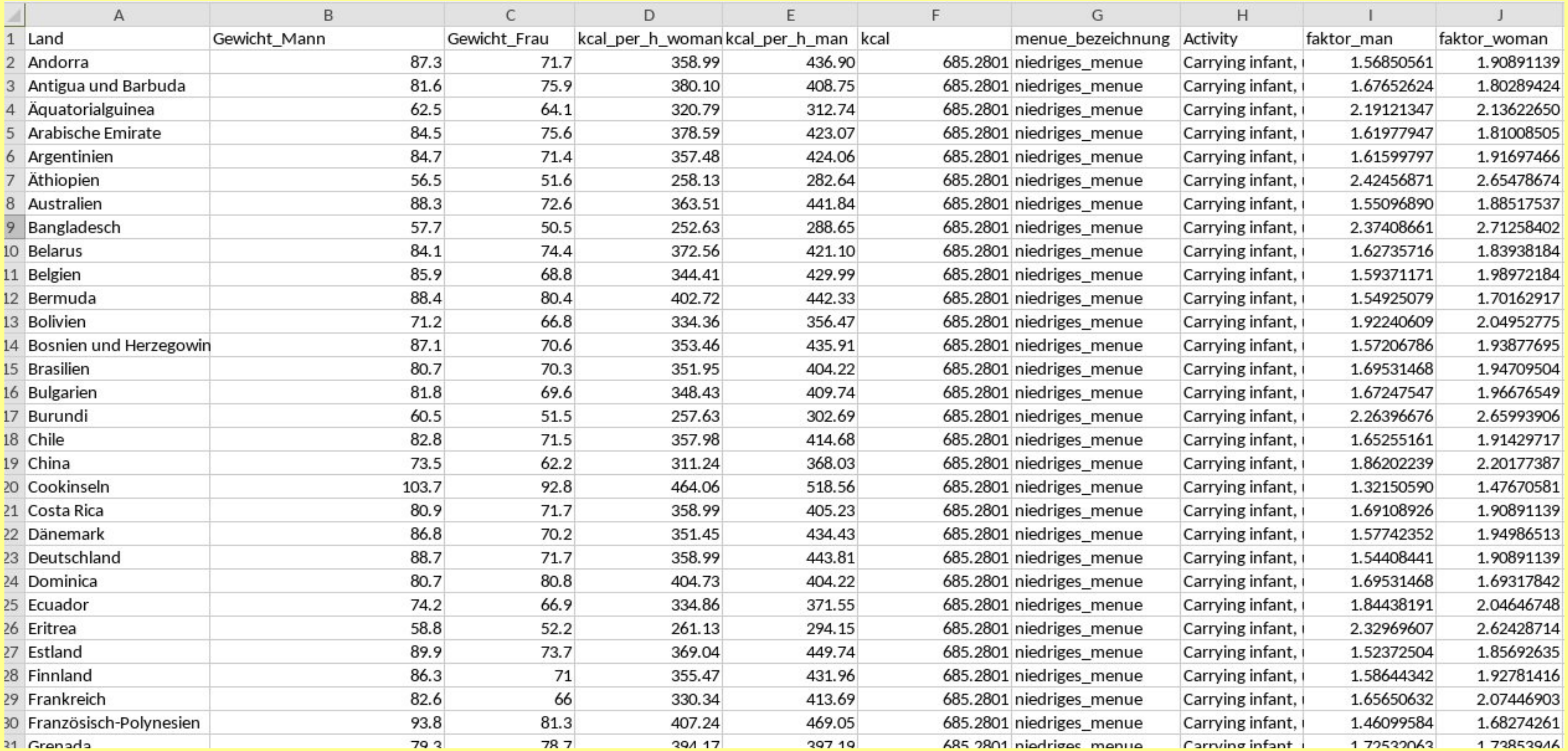
- generating Mac menues out of the Macdonalds items Dataset and calculating Calories

problem : too much records

- generating new Attributes like the factor attribute for every country's mens and womens.

Result Dataset

Land, Gewicht_Mann, Gewicht_Frau, kcal_per_h_woman, kcal_per_h_man,kcal, Activity
menue_bezeichnung : niedrig, mittel, hoch
faktor_man, faktor_woman: A factor that indicates how many hours of training are required to burn off the menu.



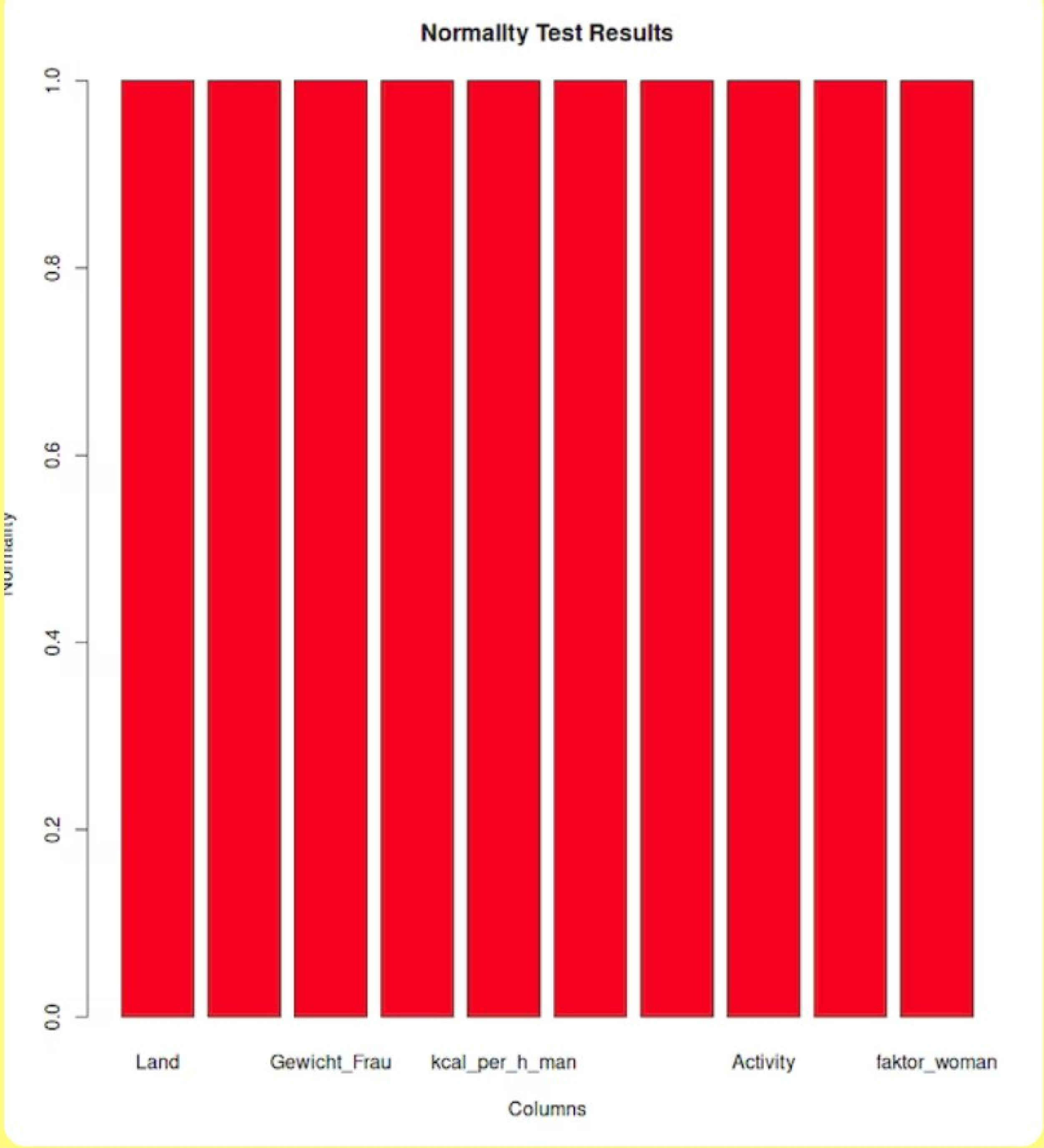
	A	B	C	D	E	F	G	H	I	J
1	Land	Gewicht_Mann	Gewicht_Frau	kcal_per_h_woman	kcal_per_h_man	kcal	menue_bezeichnung	Activity	faktor_man	faktor_woman
2	Andorra	87.3	71.7	358.99	436.90	685.2801	niedriges_menue	Carrying infant,	1.56850561	1.90891139
3	Antigua und Barbuda	81.6	75.9	380.10	408.75	685.2801	niedriges_menue	Carrying infant,	1.67652624	1.80289424
4	Äquatorialguinea	62.5	64.1	320.79	312.74	685.2801	niedriges_menue	Carrying infant,	2.19121347	2.13622650
5	Arabische Emirate	84.5	75.6	378.59	423.07	685.2801	niedriges_menue	Carrying infant,	1.61977947	1.81008505
6	Argentinien	84.7	71.4	357.48	424.06	685.2801	niedriges_menue	Carrying infant,	1.61599797	1.91697466
7	Äthiopien	56.5	51.6	258.13	282.64	685.2801	niedriges_menue	Carrying infant,	2.42456871	2.65478674
8	Australien	88.3	72.6	363.51	441.84	685.2801	niedriges_menue	Carrying infant,	1.55096890	1.88517537
9	Bangladesch	57.7	50.5	252.63	288.65	685.2801	niedriges_menue	Carrying infant,	2.37408661	2.71258402
10	Belarus	84.1	74.4	372.56	421.10	685.2801	niedriges_menue	Carrying infant,	1.62735716	1.83938184
11	Belgien	85.9	68.8	344.41	429.99	685.2801	niedriges_menue	Carrying infant,	1.59371171	1.98972184
12	Bermuda	88.4	80.4	402.72	442.33	685.2801	niedriges_menue	Carrying infant,	1.54925079	1.70162917
13	Bolivien	71.2	66.8	334.36	356.47	685.2801	niedriges_menue	Carrying infant,	1.92240609	2.04952775
14	Bosnien und Herzegowin	87.1	70.6	353.46	435.91	685.2801	niedriges_menue	Carrying infant,	1.57206786	1.93877695
15	Brasilien	80.7	70.3	351.95	404.22	685.2801	niedriges_menue	Carrying infant,	1.69531468	1.94709504
16	Bulgarien	81.8	69.6	348.43	409.74	685.2801	niedriges_menue	Carrying infant,	1.67247547	1.96676549
17	Burundi	60.5	51.5	257.63	302.69	685.2801	niedriges_menue	Carrying infant,	2.26396676	2.65993906
18	Chile	82.8	71.5	357.98	414.68	685.2801	niedriges_menue	Carrying infant,	1.65255161	1.91429717
19	China	73.5	62.2	311.24	368.03	685.2801	niedriges_menue	Carrying infant,	1.86202239	2.20177387
20	Cookinseln	103.7	92.8	464.06	518.56	685.2801	niedriges_menue	Carrying infant,	1.32150590	1.47670581
21	Costa Rica	80.9	71.7	358.99	405.23	685.2801	niedriges_menue	Carrying infant,	1.69108926	1.90891139
22	Dänemark	86.8	70.2	351.45	434.43	685.2801	niedriges_menue	Carrying infant,	1.57742352	1.94986513
23	Deutschland	88.7	71.7	358.99	443.81	685.2801	niedriges_menue	Carrying infant,	1.54408441	1.90891139
24	Dominica	80.7	80.8	404.73	404.22	685.2801	niedriges_menue	Carrying infant,	1.69531468	1.69317842
25	Ecuador	74.2	66.9	334.86	371.55	685.2801	niedriges_menue	Carrying infant,	1.84438191	2.04646748
26	Eritrea	58.8	52.2	261.13	294.15	685.2801	niedriges_menue	Carrying infant,	2.32969607	2.62428714
27	Estland	89.9	73.7	369.04	449.74	685.2801	niedriges_menue	Carrying infant,	1.52372504	1.85692635
28	Finnland	86.3	71	355.47	431.96	685.2801	niedriges_menue	Carrying infant,	1.58644342	1.92781416
29	Frankreich	82.6	66	330.34	413.69	685.2801	niedriges_menue	Carrying infant,	1.65650632	2.07446903
30	Französisch-Polynesien	93.8	81.3	407.24	469.05	685.2801	niedriges_menue	Carrying infant,	1.46099584	1.68274261
31	Grenada	79.3	78.7	394.17	397.19	685.2801	niedriges_menue	Carrying infant,	1.72532063	1.73853946

Showcase

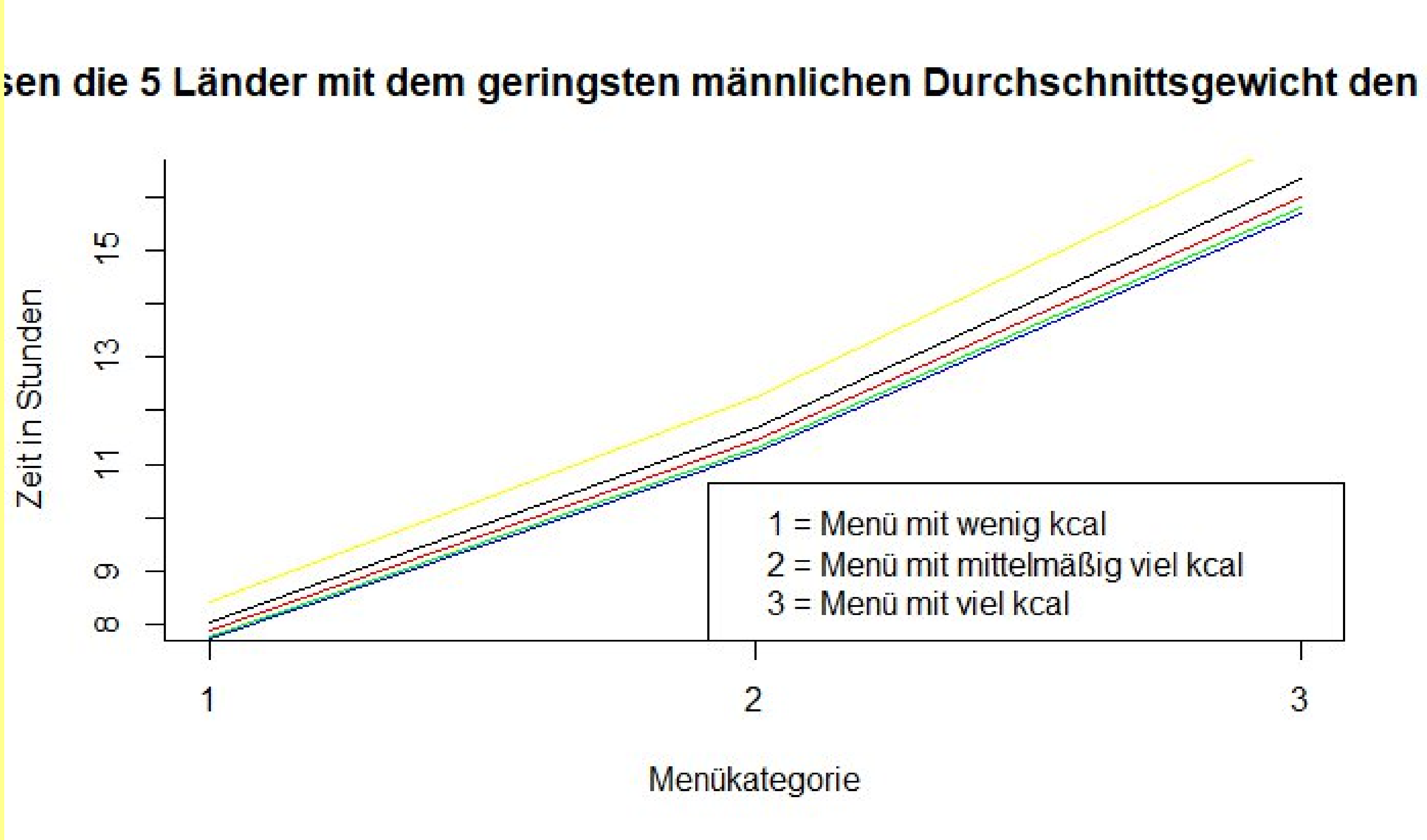
Visualisation done using R

Distribution of data

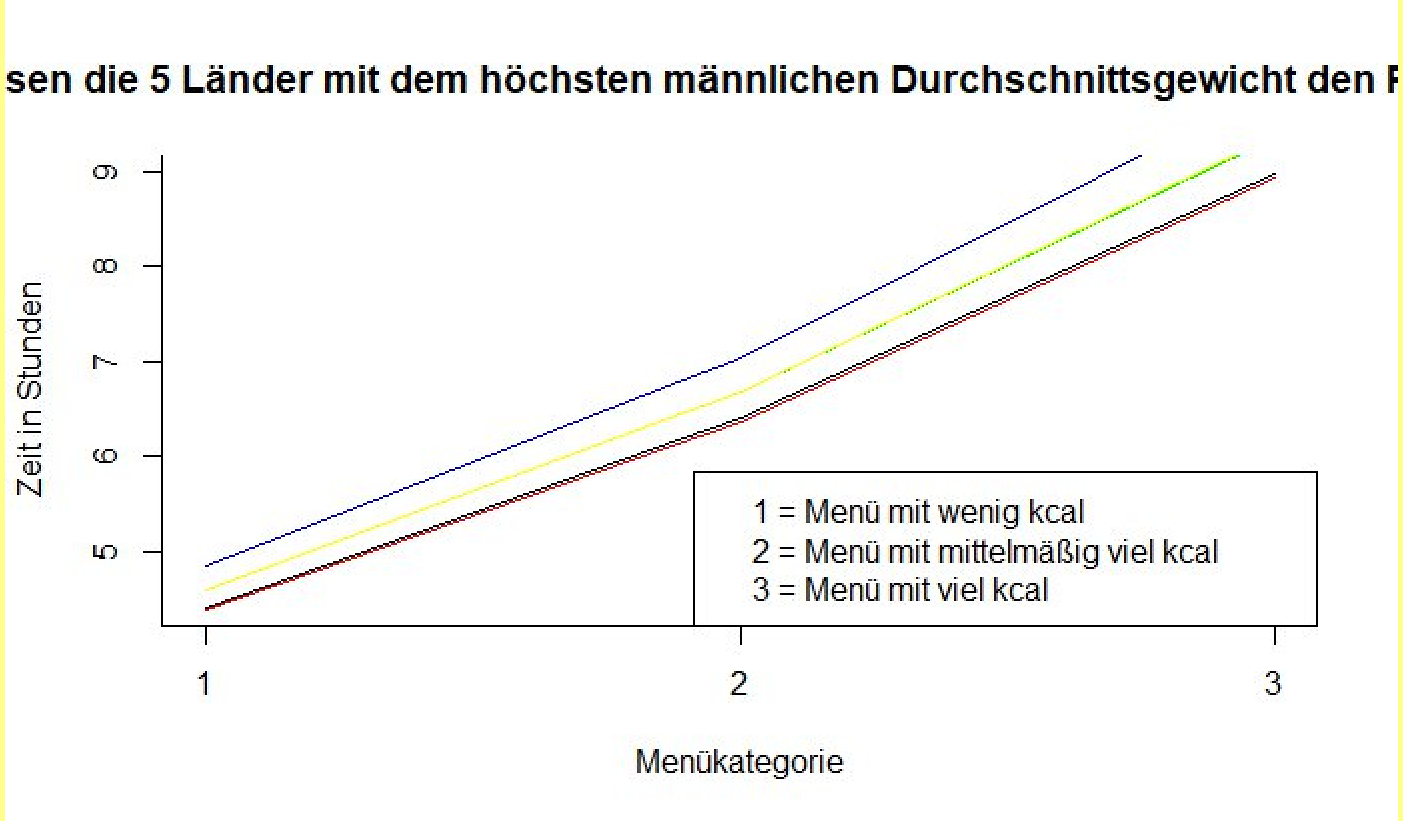
Using the Shapiro-Wilk Test it was found that none of the attributes were normally distributed.



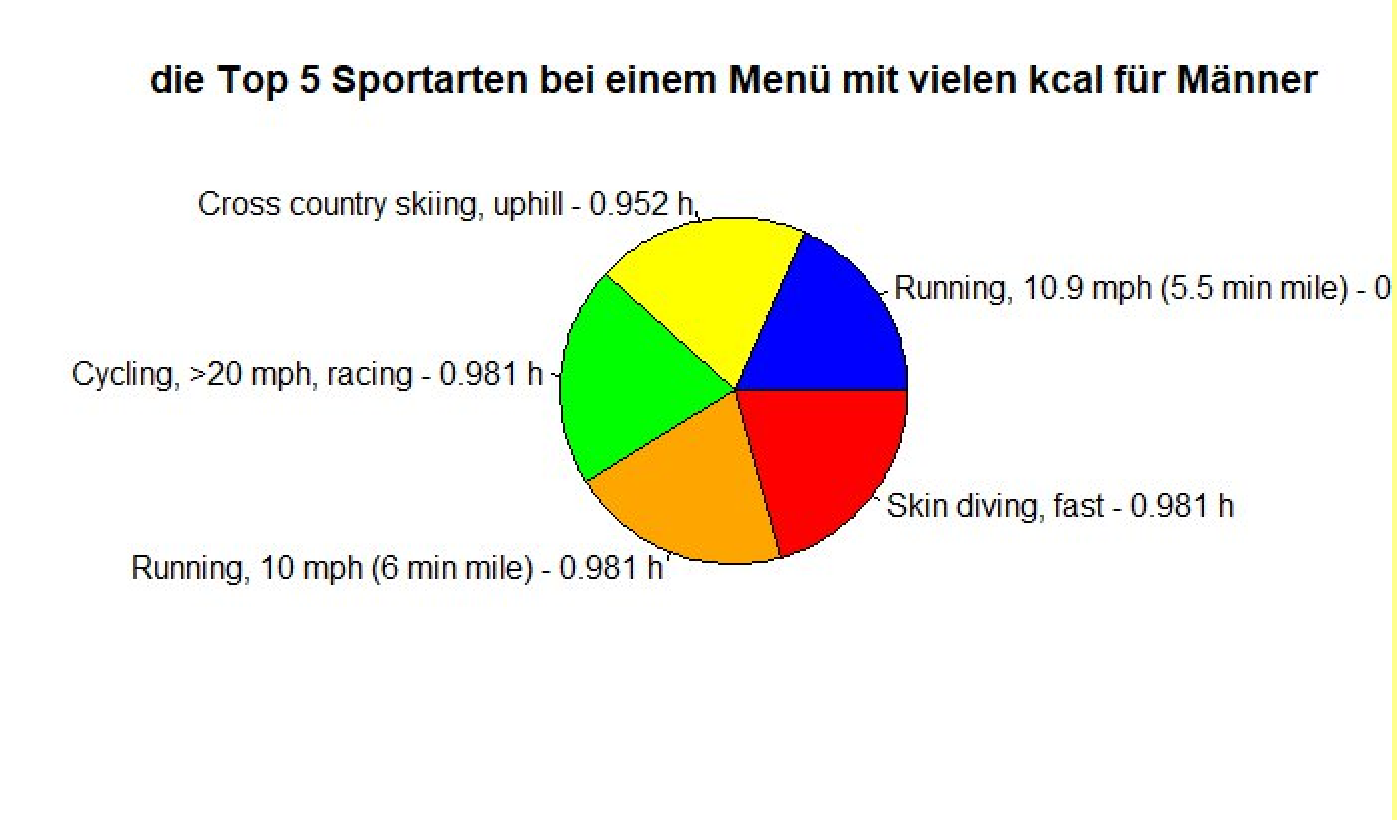
Top 5 Länder mit geringstem Durchschnittsgewicht und Dauer des Trainings



Top 5 Länder mit höchstem männlichen Durchschnittsgewicht und Dauer für Verbrauch



Top 5 Sportarten bei Menü mit vielen Kalorien für Deutschland neu


die Top 5 Sportarten bei einem Menü mit vielen kcal für Männer
Cross country skiing, uphill - 0.952 h
Running, 10.9 mph (5.5 min mile) - 0
Cycling, >20 mph, racing - 0.981 h
Skin diving, fast - 0.981 h
Running, 10 mph (6 min mile) - 0.981 h