Praktikum Data Mining

Energieverbrauch und CO2-Emmisionen
Vorhersage und Clustering auf Finanzdaten

Oliver Fesseler Maria Florusß Stefan Seibert
Daniel Grießhaber

8. Mai 2014

Inhaltsverzeichnis

Energieverbrauch und CO_2-Emmission

Datenverwaltung und Statistik

Einlesen der Daten, Hinzufügen der GPS Koordinaten, Abspeichern in neuer Datei

Bei der Umsetzung der Aufgaben haben wir mit verschiedenen Darstellungsformen der Daten experimentiert. Im ersten Plot fallen vor allem die Vielverbraucher einzelner Energieformen auf. Beim zweiten Plot können die verschiedenen Energiemixe pro Land direkt miteinander verglichen werden, da sie alle im selben Maßstab nebeneinander dargestellt werden.

Ausgehend von der implementierten Visualisierung des Energieverbrauchs der Länder: Nennen Sie die 3 Ihrer Meinung nach interessantesten Beobachtungen.

1. Durch die wenigen Industrieländer mit signifikant höherem Energieverbrauch, wie China oder die USA, wird der Plot so verzerrt, dass die Länder mit durchschnittlichem Verbrauch im Plot so gestaucht werden, dass sie kaum zu erkennen sind. Dies könnte behoben werden, wenn die Daten mit den Einwohnerzahlen aller Länder normalisiert werden würden. So könnte der Pro-Kopf-Verbrauch berechnet werden, was einen besseren Vergleich der einzelnen Länder bietet.

2. Dieses Prinzip wird klar bei der Betrachtung von China und Indien, die von der Einwohnerzahl her vergleichbar sind (China 1,3 Mrd., Indien 1,2 Mrd.[1]). China zeigt einen weitaus höheren Verbrauch als Indien an den in beiden Ländern häufigen Energieformen Kohle und Öl. Zusätzlich wären noch andere Normalisierungsfaktoren interessant:

    - Bruttoinlandsprodukt
    - Außenhandelsstatistik oder Export der Länder in US$

[1] Stand 2012, Quelle: Wikipedia

- Technologieindex

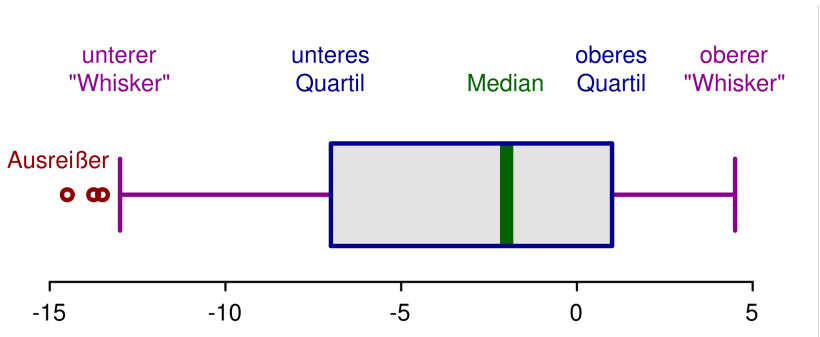
3. Die zwei Länder mit dem höchsten Energieverbrauch sind China und die USA. Dies fällt bei der Betrachtung des Gesamtenenergieverbrauchs auf. Bei der Betrachtung des Verbrauchs einzelner Energieformen wirkt es als sei China durch seinen hohen Kohleverbrauch weit vor den USA.

Abgabe: Relevante Dateien

- `energy_consumption_per_country.py` und `energy_consumption_per_country_V2.py`
  - Implementierung Aufgabe 2.1.2: 1) - 2)
- `energy_consumption_per_country.pdf`
  - Ausgabe des Skripts `energy_consumption_per_country.py`
- `energy_consumption_per_country_V2.pdf`
  - Ausgabe des Skripts `energy_consumption_per_country_V2.py`
- `appendGeoCoordinates.py`
  - Implementierung Aufgabe 2.1.2: 3) - 5)
- `EnergyMixGeo.csv`
  - Ausgabe des Skripts `appendGeoCoordinates.py`

Statistik der Daten

Erklären Sie sämtliche Elemente eines Boxplot (allgemein).



- Ausreißer (“Outliers“), in Python mit ’sym’ zu setzen.
  Daten die außerhalb der Whisker liegen und somit als Ausreißer deklariert werden.

- Whisker, Länge in Python mit 'whis' zu setzen.
  Standartmäßig die 1,5-fache Länge des entsprechenden Quartils.
- Quartil
  Die Quartile sind Bestandteile der Box, welche 50 % aller Daten enthält. Dabei enthält das obere Quartil, die 25%, die über dem Median, das untere Quartil, die 25%, die unter dem Median liegen.
- Median
  Der mittlere Wert (nicht Mittelwert oder Durchschnitt), der aus dem gesamten Datensatz ermittelt wird. Er teilt den Boxplot in zwei Hälften, die wiederum jeweils in Whisker und Quartil unterteilt werden.

Diskutieren Sie die im Boxplot angezeigte Statistik der Energieverbrauchdaten.

Abgabe: Relevante Dateien

-