DSML Project Paper

Semesterarbeit DSML - SS 2022

Studenten	Iris Lüthi, Maja Velkova, Yannik Zimmermann
Dozenten	Prof. Dr. M. Krebs, Prof. Dr. P. Collovà
Modul	DSML
Semester	6. Semester
Datum	10. Juni 2022

Inhaltsverzeichnis

1 Einleitung

Im Rahmen des Moduls Data Science and Machine Learning DSML werden die Studierenden der Berner Fachhochschule dazu angehalten eine empirische Arbeit auf dem Gebiet der Datenanalyse oder des maschinellen Lernens unter Verwendung von R oder Python zu erarbeiten.

Das Autorenteam hat sich für vier verschiedene Datensets mit ähnlichem Kontext entschieden. In der folgenden Arbeit will das Autorenteam mithilfe der Datensets herausfinden, ob es einen Zusammenhang zwischen der weltweiten Temperaturentwicklung und der Luftzusammensetzung von verschiedenen Gasen gibt. Hierfür werden erst die weltweiten Veränderungen der Temperaturen analysiert, worauf eine genauere Analyse Temperatur, Luftqualität und Gas-Zusammensetzung der USA durchgeführt wird. Auf die Analyse der Datensets folgt die Anwendung von ML-Algorithmen auf die Daten, um eine Vorhersage für die Luftqualität und die Temperaturentwicklung zu treffen. Schlussendlich werden die Ergebnisse von dem Autorenteam zusammengefasst, woraufhin ein persönliches Fazit folgt.

2 Data Source und Datasets

Alle Datensets, die im Rahmen dieser Semesterarbeit genutzt werden, wurden auf der Plattform Kaggle veröffentlicht. Die Plattform bietet über 50'000 öffentliche Datensets. Diese können dabei direkt in Online-Notebooks analysiert werden. Kaggle bietet nebenbei einige Kurse zur Datenanalyse mithilfe verschiedenster Tools und Programmiersprachen.

2.1 Dataset "Daily Temperature of Major Cities"

Das "Daily Temperature of Major Cities" Dataset besteht aus acht Spalten und 2,9 Millionen Zeilen. In der originalen Datenstruktur sind folgende Spalten enthalten:

- Region
- Country
- State
- City
- Day
- Month
- Year
- AvgTemperature

Mithilfe des folgenden Befehls lässt sich herausfinden, wie viele Zeilen identisch sind. In diesem Dataset sind insgesamt 20'715 identisch und müssen daher vor dem nächsten Schritt gefiltert werden.

```
sum(duplicated(raw_city_temps))
```

Bei der Untersuchung des Datasets stellte sich heraus, dass sich einige der Temperaturwerte unterhalb von -90° Fahrenheit befinden. Aufgrund der unrealistischen Werte werden diese Zeilen für die weitere Analyse entfernt. Des Weiteren befinden sich mögliche Tippfehler in der Datenerhebung, da Temperaturwerte für die Jahre 200 und 201 erfasst wurden. Diese Werte werden ebenfalls gefiltert. Zuletzt wurden die Werte aus dem Jahr 2020 entfernt, da die Datenerfassung für dieses Jahr nicht abgeschlossen wurde. Die Bereinigung wurde durch folgenden Befehl vorgenommen:

```
# Data preprocessing for further analysis
city_temps = raw_city_temps %>%
  # remove duplicates
  distinct() %>%
  # remove temp values below -50f because they seem like default/null values
  filter(AvgTemperature > -50) %>%
  # remove values where year is below 1950 because those are probably typos (200, etc)
  filter(Year > 1950) %>%
  # remove year 2020 because of really low number of observations
  filter(Year != 2020)
```

Die Durchschnittstemperatur wird in Fahrenheit angegeben. Für ein besseres Verständnis wurde die Spalte “AvgTemperatureInCelcius” durch folgenden Befehl hinzugefügt:

```
city_temps['AvgTemperatureInCelcius'] = fahrenheit.to.celsius(city_temps$AvgTemperature)
```

2.2 Dataset “Global Air Quality Index (AQI)”

2.3 Dataset “US Pollution”

3 Analyse

Txxt

```
city_temps %>%
  group_by(Year) %>%
  summarize_at(vars(AvgTemperatureInCelsius), list(AvgTemp=mean)) %>%
  ggplot(aes(Year,AvgTemp)) +
  geom_smooth(method = 'loess', formula = 'y ~ x') +
  geom_line() +
  labs(
    x="Jahr",
    y="Durchschnittstemperatur in Celsius",
    title = "Gemessene Durchschnittstemperatur in Grossstädten rund um die Erde"
  )
```

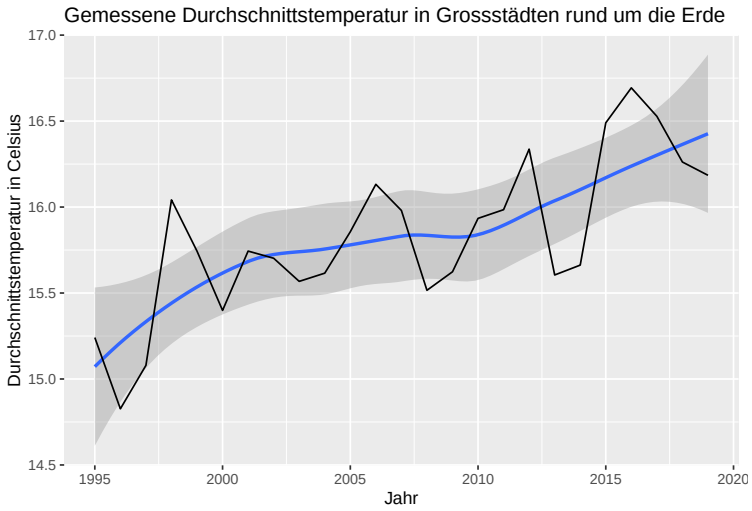

Abbildung 1: Gemessene Durchschnittstemperatur in Grossstädten rund um die Erde

xtxt

4 Resultate und Diskussion

5 Fazit

Literaturverzeichnis

Selbstständigkeitserklärung

Die Länge des vorliegenden Textes ab und inklusive Kapitelüberschrift 1 bis vor diesen Abschnitt beträgt XXXX Wörter.

Wir bestätigen, die vorliegende Arbeit selbständig verfasst zu haben. Sämtliche Textstellen, die nicht von uns stammen, sind als Zitate gekennzeichnet und mit dem genauen Hinweis auf ihre Herkunft versehen. Die verwendeten Quellen (gilt auch für Abbildungen, Grafiken u.ä.) sind im Literatur- bzw. Quellenverzeichnis aufgeführt.

Bern, 10.06.2022

Iris Lüthi

Maja Velkova

Yannik Zimmermann