Semesterarbeit des Moduls  
Data Collection, Integration and Pre-processing (CIP02)

ETL-Projektdokumentation

zum Thema  
Entwicklung von Sportgrossanlässen und Bruttoinlandprodukt sowie deren Korrelation von 1950-2019

Hochschule Luzern HSLU

Studiengang: MSc Applied Information and Data Science

Studenten: Régis Andréoli, Micha Käser  
Einreichdatum: 19.05.2021

Gruppennummer: 50

**Inhaltsverzeichnis**

[1 Ausgangslage 3](#_Toc71039712)

[2 Datenquellen 3](#_Toc71039713)

[3 Fragestellungen 4](#_Toc71039714)

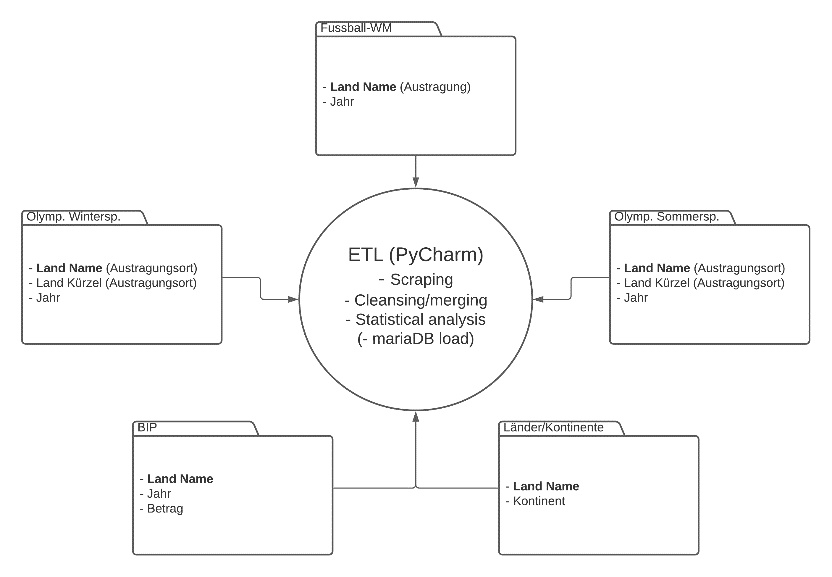
[4 Tools 4](#_Toc71039715)

[5 ETL-Prozess 4](#_Toc71039716)

[6 Arbeitsteilung 4](#_Toc71039717)

# Ausgangslage

Während des Masterstudiengangs Applied Information and Data Science erarbeiten die Studierenden innerhalb des Moduls CIP02 eine Semesterarbeit, welche die erlernten Extract-, Transform- und Load-Methoden der Vorlesung in die Praxis umsetzt. Ursprünglich zu dritt gestartet, brach ein Mitstudent das Studium ab – der knappen Zeit geschuldet setzen Régis Andréoli und Micha Käser das Projekt nun zu zweit um.

Als grosse Sportenthusiasten lag es nahe, ein entsprechendes Thema zu wählen. Damit auch eine gewisse Zeitachse (Time Series) gewährleistet ist, fiel die Wahl schlussendlich auf Sportgrossanlässe (Olympiaden/Fussballweltmeisterschaften) und deren Auswirkungen auf das Bruttoinlandprodukt (BIP) der jeweiligen austragenden Nation. Solche Anlässe sind gut dokumentiert und lassen sich viele Jahre zurückverfolgen, auch das BIP als wichtige wirtschaftliche Kennzahl lässt sich bis zur Industrialisierung nachvollziehen. Damit die Resultate verlässlich und aussagekräftig sind, werden die Daten allerdings erst ab der Nachkriegszeit (1950) bis 2019 ausgewertet.

# Datenquellen

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Daten** | **Attribute** | **Link** |
| A1[[1]](#footnote-1) | BIP nach Land [[2]](#footnote-2) 1950-2019 | * Land Name * Jahr * Betrag | <https://febpwt.webhosting.rug.nl/Dmn/AggregateXs/PivotShow> |
| B1 | Fussball-Weltmeisterschaften 1930-2018 | * Land Name Austragungsort * Jahr | <https://www.fussball-wm-total.de/History/histehre.html> |
| B2 | Olympische Winterspiele 1924-2018 | * Land Name  Austragungsort * Land Kürzel * Jahr | <https://www.taschenhirn.de/sport/olympische-winterspiele/> |
| B3 | Olympische Sommerspiele 1896-2016 | * Land Name  Austragungsort * Land Kürzel * Jahr | <https://www.taschenhirn.de/sport/olympische-sommerspiele/> |
| C1 | Länder und Kontinente | * Land Name * Kontinent | [https://www.bernhard-gaul.de/wissen/staatenerde.php#uebneu](https://www.bernhard-gaul.de/wissen/staatenerde.php%23uebneu) |
| C2 | Ländercodes | * Land Name * Land Kürzel | https://www.laenderdaten.de/kuerzel/iso\_3166-1.aspx |

Die Tabellen der Sportgrossanlässe werden in einem Dataset zusammengefasst und nach Jahr sortiert. Die Korrelationsberechnung zum BIP werden auf Basis Jahr und Land durchgeführt.

# Fragestellungen

1. Welches Land war zwischen 1950-2019 am häufigsten Gastgeber eines Sportgrossanlasses (Sommer- und Winterolympiade, Fussballweltmeisterschaft)?
2. Welcher Kontinent verzeichnet den grössten BIP-Zuwachs in der Periode von 2000-2019?
3. In welcher Dekade wurden auf welchem Kontinent die meisten Sportgrossanlässe (Sommer- und Winterolympiade, Fussballweltmeisterschaft) durchgeführt?
4. Gibt es eine statistisch feststellbare, signifikante Korrelation zwischen der Durchführung von Sportgrossanlässen und der Veränderung des BIPs der Gastgebernation?

# Tools

|  |  |
| --- | --- |
| **Zweck** | **Tools** |
| Repository Code | GitHub |
| Repository Daten | GitHub/Switch Drive |
| Kommunikation | Slack, Zoom |
| Entwicklungsumgebung | Ubuntu (VM oder nativ), PyCharm, mariaDB, Microsoft Excel/LibreOffice Calc (manuelle Verunreinigung Daten) |
| Diagramme | Lucidchart |

# ETL-Prozess

Nach dem Scrapen der Daten müssen diese verunreinigt werden, da sie bereits eine sehr hohe Qualität aufweisen. Dafür werden in Microsoft Excel oder LibreOffice Calc BIP-Beträge in verschiedene Währungen umgerechnet, Strings mit Zahlen angereichert und unterschiedliche Sprachen verwendet.

Im nächsten Schritt werden diese Verunreinigungen mit PyCharm korrigiert bzw. bereinigt. Die sauberen Tabellen werden gemerged und die Fragestellungen in PyCharm ausgewertet und ausgerechnet. Zusätzlich sollen die sauberen Datasets in eine mariaDB-Datenbank eingelesen werden.

# Arbeitsteilung

|  |  |  |
| --- | --- | --- |
| **Wer** | **Datenquelle** | **Was** |
| Régis Andréoli | A1, B1, C1 | Datenquelle scrapen (B1, C1), CSV exportieren (A1) und Daten mit Python/Pandas bereinigen/anreichern |
| Micha Käser | B2, B3 | Datenquelle scrapen (B2, B3) und Daten mit Python/Pandas bereinigen/anreichern |
| Team | Alle | Daten mergen, joinen und gruppieren. Fragestellungen beantworten. Danach mariaDB aufsetzen und bereinigte Datasets einlesen |

# Lösungsschritte

## Extract

## Transform

## Load

# Fragestellungen

# Reflexion

1. CSV-Export [↑](#footnote-ref-1)
2. Tabelle unter Quickstart Query «RGDPNA (1950-2019)» [↑](#footnote-ref-2)