|  |  |
| --- | --- |
| *Alexandra Speer*  *Nicolas Henzel*  *Hochschule der Medien*  *Sommersemester 23* |  |
|  |  |

Data Warehouse Workshop  
Prof. Dr. Klaus Freyburger

Inhaltsverzeichnis

[Daten 2](#_Toc131179257)

[Konzeptionelles Modell 2](#_Toc131179258)

[Design der Umsetzung 3](#_Toc131179259)

[Logisches Modell 3](#_Toc131179260)

[Data Wrangling 3](#_Toc131179261)

[Beispielhafte Auswertungen 4](#_Toc131179262)

[Implementierung 4](#_Toc131179263)

[Quellen 6](#_Toc131179264)

# Daten

Die Datensätze stammen von Kaggle[[1]](#footnote-1). Inhalt sind die Weltmeisterschaften der Formel 1 im Zeitraum von 1950 bis 2023. Die Daten sind durch die Ergast Developer API[[2]](#footnote-2) zusammengestellt worden, die Renndaten für nicht kommerzielle Zwecke zur Verfügung stellt.

Die Formel 1 (oft auch F1) ist eine Formelserie, die durch den Automobil Dachverband Fédération Internationale de l’Automobile (FIA) autorisiert ist. Formelserie bedeutet hierbei, dass bestimmte Regeln (Formeln) auf technischer Ebene für die Leistungsfähigkeit der Fahrzeuge festgelegt wurden, um einen Wettkampf unter gleichbleibenden Bedingungen zu ermöglichen. Die Formel 1 Weltmeisterschaft fand erstmals 1950 statt und besteht pro Saison aus bis zu 23 Grand Prix (französisch für „Großer Preis) Rennen. Dies sind Einzelrennen auf ausgewählten Rennstrecken in jeweils unterschiedlichen Ländern. Dabei sammeln die Fahrer abhängig von ihrer Endposition bei diesen Rennen Punkte. Am Ende der Saison gewinnt der Fahrer mit den meisten Punkten.  
Außerdem erhalten die Konstrukteure der Wagen Punkte, die ebenfalls am Ende der Saison ausgewertet werden.[[3]](#footnote-3)

Die Daten beinhalten alle Informationen der Formel 1 Weltmeisterschaften seit 1950:  
Rennen, Fahrer, Konstrukteure, Qualifizierungen, Rennstrecken, Rundenzeiten, Boxenstopps und Ergebnisse. Diese Informationen sind in 14 Datensätze aufgeteilt. Eine Übersicht der Daten in Python findet sich im DataExploration.ipynb.  
Die wichtigsten Erkenntnisse aus dieser Übersicht sind:

* Zellen, bei denen kein sinnvoller Wert eingetragen ist, sind mit \N markiert.  
  Das wird im Data Wrangling behoben
* Die Daten sind bereits „tidy“:  
  Jede Zeile beinhaltet eine Beobachtung,   
  jede Spalte beschreibt eine Variable,   
  in jeder Zelle ist genau ein Wert (und nicht mehrere).[[4]](#footnote-4)
* Die Datensätze besitzen mindestens eine Id Spalte (Primärschlüssel). wodurch sich die Daten leicht referenzieren lassen. Manche Datensätze haben weitere Id Spalten aufgelistet, die als Fremdschlüssel verwendet werden können.

# Konzeptionelles Modell

# Design der Umsetzung

## Logisches Modell

## Data Wrangling

Die Daten wurden im ersten Schritt im DataExploration.ipynb untersucht und in einer flachen Form zusammengeführt. Dabei haben wir uns auf die Fahrer Daten fokussiert und die Id Spalten genutzt, um die unterschiedlichen Datensätze miteinander zu joinen. Die Konstrukteur Daten sind im ER-Modell enthalten. Die Struktur dient zur ersten Übersicht, sowie beispielhaften Verbindung der Datensätze zu einem finalen Datensatz. Damit wurden zusätzlich Auswertungen beispielhaft im Python Skript umgesetzt.

Die Datensätze, die genutzt wurden, sind:

* circuits.csv
* driver\_standings.csv
* drivers.csv
* lap\_times.csv
* pit\_stops.csv
* results.csv
* races.csv
* status.csv
* constructors.csv
* constructor\_standings.csv
* constructor\_results.csv
* qualifying.csv
* sprint\_results.csv

Der seasons.csv Datensatz wurde nicht genutzt, da hier keine neuen Informationen enthalten sind.

Zum Laden der Daten in DWC wurde eine JSON Datei erstellt (180\_FormulaOne.json), die die Tabellen und Verbindungen anlegt.  
Folgende Anpassungen haben wir vorgenommen:

* Driver\_standings.csv - positionText Spalte entfernt
* Lap\_times.csv - Neuen Datensatz lap\_times\_grouped.csv erstellt, der gruppierte Werte enthält

# group laptimes of drivers for each different race by taking the mean over all laps of a driver for specific race

# milliseconds could be converted into minutes:seconds.milliseconds format if needed

df\_grouped\_lap\_times = df\_lap\_times.groupby(

['driverId','raceId']

)['milliseconds'].mean().reset\_index()

# rename column for more clarity

df\_grouped\_lap\_times.rename(

{'milliseconds':'avgMillisecondsLap'},axis=1, inplace= True

)

df\_grouped\_lap\_times.to\_csv("Data/lap\_times\_grouped.csv", index=True)

* Pit\_stops.csv - duration Spalte entfernt
* Results.csv - positon, positionText Spalte entfernt
* Constructor\_standings.csv - positionText Spalte entfernt
* Sprint\_results.csv - positon, positionText Spalte entfernt
* Alle „URL“ Spalten in jedem Datensatz entfernt (für die Auswertung nicht relevant)
* Alle Spalten mit Datum auf das Format YYYY-MM-DD angepasst
* Werte in allen mit \N markierten Zellen werden entfernt, damit beim import in DWC korrekt NULL eingetragen werden kann

Die konkreten Schritte sind im DataWrangling.ipynb nachzuvollziehen.   
  
Die beispielhaften Auswertungen werden im Anschluss erläutert.

## Beispielhafte Auswertungen

Vorschläge zur Auswertung finden sich in der DataExploration.ipynb Datei.  
Wir wollen damit folgende Fragen beantworten:

* Wie ist das Verhältnis zwischen Start und Endposition?
* Welcher Hersteller hat die meisten Weltmeisterschaften gewonnen?
* Welcher Fahrer hat die meisten Weltmeisterschaften gewonnen?
* In welcher Saison sind die meisten Unfälle passiert?
* Welcher Hersteller hat die meisten Defekte / Ausfälle / Unfälle?
* Welche Strecken werden mit den kürzesten, welche mit den längsten Zeiten gefahren?

# Implementierung

Für die Implementierung haben wir eine JSON Datei (180\_FormulaOne.json) erstellt, wodurch das E/R Modell per import in DWC angelegt wurde.

Die erste Herausforderung war, dass die Datumsformate in der richtigen Form sein mussten, wir haben YYYY-MM-DD gewählt. Spalten, die einen Zeitwert enthalten haben, wurden nicht korrekt erkannt, da diese nicht im H:MM:SS Format vorliegen, sondern H:MM:SS.MS.

 Aus diesem Grund haben wir diese Spalten als String importiert, nachdem die Fehlermeldung in DWC aufgetreten ist.  
Außerdem haben wir bemerkt, dass einige Zeichenketten länger waren als vermutet, wodurch die Tabellen händisch angepasst werden mussten (z.B. surname Spalte in Drivers Tabelle, kommt mit 20 Zeichen nicht aus, weshalb die Länge auf 40 Zeichen erhöht wurde). Zusätzlich gibt es Spalten mit Zahlen, die auf den ersten Blick wie Integers aussehen, nach dem import jedoch Float sind, was wir auch anpassen mussten (z.B. points Spalte in results.csv).

Unser E/R Modell in DWC sieht folgendermaßen aus:

Diagram, schematic

Description automatically generated

# Quellen

<https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020>  
Zuletzt besucht am 30.03.2023 19 Uhr

<http://ergast.com/mrd/>  
Zuletzt besucht am 30.03.2023 19 Uhr

<https://de.wikipedia.org/wiki/Formel_1>  
Zuletzt besucht am 30.03.2023 19 Uhr

<https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html>  
Zuletzt besucht am 30.03.2023 19 Uhr

1. <https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020> [↑](#footnote-ref-1)
2. <http://ergast.com/mrd/> [↑](#footnote-ref-2)
3. <https://de.wikipedia.org/wiki/Formel_1>  
    [↑](#footnote-ref-3)
4. https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html [↑](#footnote-ref-4)