

Workshop zu
Datenqualität

City Lab Berlin

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Workshop zu Datenqualität

City Lab Berlin

Datum (TBD)

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Einführung

Plan

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:

Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

① Theorie

- Was ist Datenqualität?
- Relevanz von Datenqualität

② Praxis

- Datenqualität für Excel-Dateien
- Datenqualität für CSV-Dateien
- Excel vs. CSV: Welches sollte ich nutzen?
- Andere Datenformate
- Häufige Datenqualitäts-Probleme

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

1. Theorie

Was ist Datenqualität?

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:

Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Wie bewertet man die Qualität eines
Datensatzes?

- 2 Dimensionen:
 - Wie sind die Daten *aufbereitet*?
 - Wie sind die Daten *bereitgestellt*?

Was ist Datenqualität?

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- Neben der Aufbereitung und Bereitstellung der Daten, kann man auch hinterfragen, wie die Daten tatsächlich erhoben wurden
 - Werden wir aber bei diesem Workshop nicht diskutieren

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

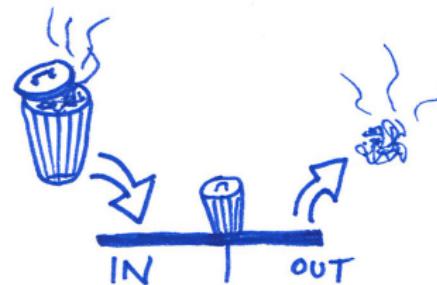
Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Warum reden wir darüber?

- Datenqualität ist so wichtig weil:
“Garbage in, garbage out”
 - Die Qualität des Inputs bestimmt die Qualität des Outputs!



Bewertung der Datenqualität

3 Beispiele von Heransgehenweisen:

- ① FAIR-Prinzipie
- ② 5-Sterne-Modell
- ③ NQDM Merkmale der Datenqualität

Bewertung der Datenqualität

① FAIR

- FAIR-Data ist:

- **F**indable (Auffindbar)
- **A**ccessible (Zugänglich)
- **I**nteroperable (Interoperabel)
- **R**eusable (Wiederverwendbar)

Quelle: Wilkinson, M. D. et al. (2016)

Bewertung der Datenqualität

- **Findable:** *Wo/wie sind die Daten auffindbar?*
 - Datensatz ist registriert oder indiziert (z.B.: in einem Datenportal)
 - Dauerhafte Bezeichnung (Stable ID)
 - Ausführliche Beschreibung des Inhalts

Bewertung der Datenqualität

- **Accessible:** *Wie kann ich auf die Daten zugreifen?*
 - Klar und kostenlosen Zugriffsprotokoll
 - *Metadaten* (Beschreibung der Daten) und *Daten* (Inhalt) sind abrufbar
 - Metadaten sind **immer** verfügbar, auch wenn die Daten nicht mehr abrufbar sind

Quelle: Wilkinson, M. D. et al. (2016)

Bewertung der Datenqualität

- **Interoperable:** *Wie leicht ist es, die Daten im Zusammenhang mit anderen Systemen oder Datensätzen zu verwenden?*

- Formale und zugängliche Sprache
- Verweise auf andere Daten
- Mögliche Integration mit anderen Datensätzen

Bewertung der Datenqualität

- **Reusable:** *Wie kann ich diese Daten wiederverwenden?*

- Detaillierter Herkunft: Wo kommen die Daten her?
- Verständliche Lizenzierung

Quelle: Wilkinson, M. D. et al. (2016)

Bewertung der Datenqualität

② 5-Sterne-Modell

Das 5-Sterne Modell ist wie folgt definiert:

- ★ Stellen Sie Daten im Web unter einer offenen Lizenz bereit, das Datenformat für die Bereitstellung ist Ihnen überlassen.
- ★★ Stellen Sie Daten in einem strukturierten Format bereit.
- ★★★ Verwenden Sie offene, nicht proprietäre Formate.
- ★★★★ Verwenden Sie URLs, um Dinge zu bezeichnen, damit Daten verlinkt werden können.
- ★★★★★ Verlinken Sie Ihre Daten mit anderen Daten, um Kontexte herzustellen.



Quelle: Herausgegeben von der Deutschen Zentrale für Tourismus e.V. (DZT).
Horster und Karle (2019): Braucht der Tourismus Open Data? In Anlehnung an Tim Berners-Lee (2015): 5★ Offene Daten.
Illustration: Lena Modrow.



Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Bewertung der Datenqualität

Daten im Web (ex.: PDF)



Daten in strukturiertem Format (ex.: XLS)



Daten in strukturiertem, nicht proprietärem
Format (ex.: CSV)

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Bewertung der Datenqualität



Verwendung von eindeutigen URLs (ex.: RDF)



Verlinkung der eigenen Daten mit anderen
Daten (ex.: LOD)

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Bewertung der Datenqualität

- ③ **Merkmale der Datenqualität** aus dem
“Normentwurf für qualitativ hochwertige
Daten und Metadaten” (NQDM,
Fraunhofer FOKUS)

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Bewertung der Datenqualität

- Liste von Merkmalen der Datenqualität:

- Aktualität
- Fehlerfreiheit
- Genauigkeit
- Konformität
- Konsistenz
- Transparenz & Vertrauenswürdigkeit
- Verlässlichkeit
- Verständlichkeit
- Vollständigkeit
- Zugänglichkeit & Verfügbarkeit

Quelle: "Leitfaden für qualitativ hochwertige Daten und Metadaten",
Fraunhofer FOKUS/NQDM, S. 14.
<https://www.nqdm-projekt.de/de/downloads/leitfaden>

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Quiz

Datensatz von Berlin Open Data

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- **Schulbausanierung Sommer 2020**
 - <https://daten.berlin.de/datensaetze/schulbausanierung-sommer-2020>

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

▶ Datensätze

Datensätze	Schulbausanierung Sommer 2020
Arbeitsmarkt	Berlin investiert in den kommenden Jahren 5,5 Milliarden Euro in die Sanierung und in den Bau von Schulgebäuden. Auch in den Sommerferien 2020 gehen an den Berliner Schulen die Bau- und Sanierungsmaßnahmen weiter.
Bildung	
Demographie	
Geographie und Stadtplanung	
Gesundheit	
Jugend	
Kunst und Kultur	
Öffentliche Verwaltung, Haushalt und Steuern	
Protokolle und Beschlüsse	
Sonstiges	

Lizenz: Creative Commons Namensnennung 

Kategorie: Bildung

Geographische Abdeckung: Berlin

Geographische Granularität: Berlin

Zeitliche Granularität: Keine

Veröffentlicht: 27.07.2020

Aktualisiert: 27.07.2020

Veröffentlichende Stelle: Senatsverwaltung für Bildung, Jugend und Familie

E-Mail Kontakt: karen.koenig AT senbjf.berlin.de

Website: <https://www.berlin.de/sen/bildung/service/daten/>

Tags: Bauen | Sanieren | Schulbaumaßnahmen ... (2 weitere)

Kommentare: 0

Fragen

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- Wie gut ist dieser Datensatz bezüglich der FAIR-Prinzipien?
 - **Findable:** *Wo finde ich die Daten?*
 - **Accessible:** *Wie kann ich auf die Daten zugreifen?*
 - **Interoperable:** *Wie leicht ist es, diese Daten zu verwenden?*
 - **Reusable:** *Weiß ich, wie ich diese Daten wiederverwenden kann/darf?*
- Wie optimal ist das Datenformat?

Analyse

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- **Findable:** *Wo finde ich die Daten?*

✓ **Datensatz registriert oder indiziert**

✗ **Dauerhafte Identifikationsnummer**

≈ **Ausführliche Beschreibung**

Analyse

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- **Accessible:** *Wie kann ich auf die Daten zugreifen?*

- ✓ Klar und kostenlosen Zugriffsprotokoll
- ✓ Metadaten (Beschreibung der Daten) und Daten (Inhalt) sind abrufbar

Analyse

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- **Interoperable:** Wie leicht ist es, diese Daten zu verwenden?

✓ Formale und zugängliche Sprache

✗ Verweise auf andere Daten

≈ Ermöglichte Integration mit anderen Datensätzen

Analyse

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- **Reusable:** Weiß ich, wie ich diese Daten wiederverwenden kann/darf?

 Detaillierter Herkunft

 Klare und zugängliche Datennutzungs-lizenz

Analyse

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- **Datenformat:**

≈ Aktuell = **Excel**



✓ Optimal = **CSV**



Zusammenfassung

Theorie

- Datenqualität umfasst viele verschiedene Aspekte
 - Nicht alle diese Aspekte könnten bei diesem Workshop angesprochen werden
- Heute: Besonderer Fokus auf der *Aufbereitung* von Datensätzen (Kontextunabhängig)
 - Insbesondere: Konformität und Verständlichkeit (NQDM)

Workshop zu
Datenqualität

City Lab Berlin

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

2. Praxis

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Excel

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Was ist eine Excel-Datei?

- Tabelle, die mit Excel bearbeitet werden kann
- Komplexes Dateiformat
- Stärker Fokus auf Human-Readability (im Vergleich zu anderen Dateiformaten)

Zu vermeiden in Excel

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- ① Überflüssige Text-Formatierung
(insbesondere verbundene Zellen, leere Reihen, usw.)
- Kann die Maschinenlesbarkeit der Tabelle beeinträchtigen
- Lieber effiziente Kategorisierung statt überarbeitete Tabellen

Zu vermeiden in Excel

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel



Institution	Preise (von/bis)	ABO**	Schüler	Studenten	Auszubildende	Bundesfreiwilligendienstleistende	Rentner
Große Sprechbühnen							
Deutsches Theater / Kammerspiele	4,00-48,00 €	40%	9,00 €	9,00 €	9,00 €	9,00 €	Nein
Volksbühne	6,00-40,00 €	25%	50%	50% Studententag 2x1	50%	50%	Nein
Maxim Gorki Theater	10,00-30,00 €	50%	8,00 €	8,00 €	8,00 €	8,00 €	Nein
Berliner Ensemble	5,00-30,00 €	6,00-18,00 €	9,00 €	9,00 €	9,00 €	9,00 €	9,00 €
Hebbel am Ufer (HAU 1, 2, 3)	10,00-30,00 €	10er Karte Tanzcard	5,00-10,00 €	5,00-10,00 €	5,00-10,00 €	5,00-10,00 €	5,00-10,00 €
Schaubühne am Lehner Platz	7,00-47,00 €	25% Tanzcard	9,00 €	9,00 €	9,00 €	9,00 €	Nein
Renaissance Theater Neue Theater Betriebs-GmbH	10,00-48,00 €	12,00-22,00 €	So-Do 6,00 €	So-Do 6,00 €	auf Anfrage	auf Anfrage	Nein
Kinder- und Jugendtheater							
Theater an der Parkaue	13,00-14,00 €	30%	9,00-10,00 € Kinder bis 12 Jahre 7 €	9,00-10,00 €	9,00-10,00 €	9,00-10,00 €	9,00-10,00 €
Grips Theater	6,00-20,00 €	Nein	3,00-12,00 €	3,00-12,00 €	3,00-12,00 €	3,00-12,00 €	Nein
Kleine und mittlere privatrechtlich organisierte Theater, Theater-Tanzgruppen (Konzeptförderung)							
Constanza Macras / Dorkypark GmbH	variiert	Tanzcard	variiert	variiert	variiert	variiert	variiert

Quelle: Open Data Portal, "Eintrittspreisregelungen öffentlich geförderter Berliner Kultureinrichtungen"

Zu vermeiden in Excel

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel



Art	Institution	Preise (von/bis)	ABO**	Schüler	Studenten
Große Sprechbühnen	Deutsches Theater / Kammerspiele	4,00-48,00 €	40%	9,00 €	9,00 €
Große Sprechbühnen	Volksbühne	6,00-40,00 €	25%	50%	50% Studententag 2x1
Große Sprechbühnen	Maxim Gorki Theater	10,00-30,00 €	50%	8,00 €	8,00 €
Große Sprechbühnen	Berliner Ensemble	5,00-30,00 €	6,00-18,00 €	9,00 €	9,00 €
Große Sprechbühnen	Hebbel am Ufer (HAU 1, 2, 3)	10,00-30,00 €	10er Karte Tanzcard	5,00-10,00 €	5,00-10,00 €
Große Sprechbühnen	Schaubühne am Lehniner Platz	7,00-47,00 €	25% Tanzcard	9,00 €	9,00 €
Große Sprechbühnen	Renaissance Theater Neue Theater Betriebs-GmbH	10,00-48,00 €	12,00-22,00 €	So-Do 6,00 €	So-Do 6,00 €

Quelle: Open Data Portal, "Eintrittspreisregelungen öffentlich geförderter
Berliner Kultureinrichtungen"

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel



Zu vermeiden in Excel

• Aktive Formeln

- Daten sollten nicht versehentlich vom Benutzer geändert werden

The screenshot shows a Microsoft Excel spreadsheet. The table has four columns: Country, Population, Average Population, and Greater than Average?. The 'Greater than Average?' column contains the formula =IF(B2>C2,TRUE,FALSE). A large red X is overlaid on the left side of the table. The formula bar at the top shows the formula =IF(B2>C2,TRUE,FALSE) with cell D2 selected.

	A	B	C	D	E	F
1	Country	Population	Average Population	Greater than Average?		
2	China	1,389,618,778	435,810,199	TRUE		
3	India	1,311,559,204	435,810,199	TRUE		
4	USA	331,883,986	435,810,199	FALSE		
5	Indonesia	264,935,824	435,810,199	FALSE		
6	Pakistan	210,797,836	435,810,199	FALSE		
7	Brazil	210,301,591	435,810,199	FALSE		
8	Nigeria	208,679,114	435,810,199	FALSE		
9	Bangladesh	161,062,905	435,810,199	FALSE		
10	Russia	141,944,641	435,810,199	FALSE		
11	Mexico	127,318,112	435,810,199	FALSE		

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Auswertung Excel-Dateien

- **Vorteil:** Bessere Darstellungsmöglichkeiten
- **Nachteil:** Können schnell nicht-maschinenlesbar werden

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

CSV

Was ist ein CSV?

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- Text-Datei
- Einfaches Dateiformat
- Werte werden mit einem Trennzeichen abgetrennt (“Comma Separated Values”)
- Stärker Fokus auf Machine-Readability

Was ist ein CSV?

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:

Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- So sieht ein CSV aus:

```
"id";"datum";"mitte";"friedrichshain_kreuzberg";"pankow";"charlottenburg_wilmersdorf";
"1";"2020-03-03";"3";"0";"0";"0";"0";"1";"1";"0";"1";"0";"0"
"4";"2020-03-04";"0";"2";"1";"0";"0";"0";"0";"0";"0";"0";"0"
"7";"2020-03-05";"4";"2";"2";"0";"0";"0";"0";"0";"0";"0";"2"
"10";"2020-03-06";"1";"0";"1";"2";"0";"0";"1";"0";"0";"0";"0"
"13";"2020-03-07";"0";"0";"0";"1";"0";"0";"0";"0";"0";"0";"0"
"16";"2020-03-08";"1";"1";"0";"7";"2";"0";"0";"0";"0";"0";"0"
"19";"2020-03-09";"0";"3";"2";"1";"3";"2";"3";"3";"0";"1";"3";"1"
"22";"2020-03-10";"2";"4";"3";"6";"2";"4";"5";"1";"4";"1";"2";"0"
"25";"2020-03-11";"9";"2";"6";"9";"1";"7";"1";"4";"0";"0";"1";"2"
"28";"2020-03-12";"7";"0";"4";"10";"0";"6";"5";"7";"3";"1";"0";"4"
"31";"2020-03-13";"7";"10";"9";"12";"1";"4";"7";"3";"3";"0";"1";"2"
"34";"2020-03-14";"7";"5";"0";"2";"1";"6";"2";"1";"0";"4";"0";"1"
"37";"2020-03-15";"8";"2";"0";"6";"2";"3";"1";"2";"0";"2";"0";"0"
```

Quelle: Open Data Portal, "COVID-19 in Berlin, Verteilung in den Bezirken - Gesamtübersicht"

Zu vermeiden in CSV

• Trennzeichen

- Trennzeichen sollten nicht an anderen Stellen verwendet werden (z.B.: beim Trennzeichen „,” sollte man “10.5” schreiben statt “10,5”)



	A	B	C	D	E
1		id altersgruppe	fallzahl	differenz	inzidenz
2	1	0-4	319	2	168,1
3	4	5-9	352	1	205
4	7	10-14	438	4	279,9
5	10	15-19	497	3	331,7
6	13	20-24	969	18	474,3
7	16	25-29	1235	14	430,1
8	19	30-39	2343	22	359,8
9	22	40-49	1701	5	351,3
10	25	50-59	1553	5	285,9

Quelle: Open Data Portal, “COVID-19 Fälle im Land Berlin, Verteilung nach Altersgruppen” (**verändert**)

Zu vermeiden in CSV

• Trennzeichen

- Trennzeichen sollten nicht an anderen Stellen verwendet werden (z.B.: beim Trennzeichen „,” sollte man “10.5” schreiben statt “10,5”)



A	B	C	D	E
1	id	altersgruppe	fallzahl	differenz
2	1	0-4	319	2
3	4	5-9	352	1
4	7	10-14	438	4
5	10	15-19	497	3
6	13	20-24	969	18
7	16	25-29	1235	14
8	19	30-39	2343	22
9	22	40-49	1701	5
10	25	50-59	1553	5

Quelle: Open Data Portal, “COVID-19 Fälle im Land Berlin, Verteilung nach Altersgruppen”

Zu vermeiden in CSV

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel



- Text-Formatierung

- Wird sowieso nicht gespeichert / angezeigt

	A	B	C	D
1	id	ssflag	pagenumberinocrdocument	authorfirstname
2	70526	2		0 Adler
3	70527	2		2 Bruno
4	70528	2		2 Felix
5	70529	2		2 Friedrich (Wolfgang)
6	70530	2		2 Georg
7	70531	2		2 Max
8	70532	2		2 Otto
9	70533	2		2 Viktor

Quelle: Open Data Portal, "Liste der verbotenen Bücher" (**verändert**)

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:

Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Zu vermeiden in CSV

- Text-Formatierung

- Wird sowieso nicht gespeichert / angezeigt



	A	B	C	D
1	id	ssflag	pagenumberinocrdocument	authorfirstname
2	70526	2		0 Adler
3	70527	2		2 Bruno
4	70528	2		2 Felix
5	70529	2		2 Friedrich (Wolfgang)
6	70530	2		2 Georg
7	70531	2		2 Max
8	70532	2		2 Otto
9	70533	2		2 Viktor

Quelle: Open Data Portal, "Liste der verbotenen Bücher"

Zu vermeiden in CSV

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:

Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- Überschriften (außer Spaltennamen),
Copyright-Hinweise, usw.



	A	B	C	D	E	F
1	# (c) Der Bundeswahlleiter, Wiesbaden 2017					
2	#					
3	# Wahl zum 19. Deutschen Bundestag (24. September 2017)					
4	# Endgültige Ergebnisse der Erst- und Zweitstimmen sowie der Vorperiode nach					
5	#					
6	Nr	Gebiet	gehört zu	Wahlberechtigte		
7				Erststimmen	Zweitstimme	
8				Endgültig	Vorperiode	Endgültig
9	1 Flensburg – S		1	228471	226944	228471
10	2 Nordfriesland		1	186568	186177	186568
11	3 Steinburg – D		1	176636	176731	176636
12	4 Rendsburg-E		1	200831	198903	200831

Quelle: Der Bundeswahlleiter, "Wahl zum 19. Deutschen Bundestag"

Zu vermeiden in CSV

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:

Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- Überschriften (außer Spaltennamen),
Copyright-Hinweise, usw.



Nr	Gebiet	gehört_zu	Wahlberechtigte_Erststimmen_Endgültig	Wahlberechtigte_Erststimmen_Vorperiode	\
1	Fleensburg – Schleswig	1	228471	226944	
2	Nordfriesland – Dithmarschen Nord	1	186568	186177	
3	Steinburg – Dithmarschen Süd	1	176636	176731	
4	Rendsburg-Eckernförde	1	200831	198903	
5	Kiel	1	204650	205243	
6	Plön – Neumünster	1	174937	174746	
7	Pinneberg	1	238533	235610	
8	Segeberg – Stormarn-Mitte	1	247296	244240	
9	Ostholstein – Stormarn-Nord	1	181522	180022	
10	Herzogtum Lauenburg – Stormarn-Süd	1	244930	241257	
11	Lübeck	1	181638	181923	
1	Schleswig-Holstein	99	2266012	2251796	

Quelle: Der Bundeswahlleiter, "Wahl zum 19. Deutschen Bundestag"

Hinweise

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- Zeichencodierung

- Empfehlung: UTF-8 (die Defaultcodierung beim Speichern eines CSV in Excel)
- Die Codierung sollte bewusst ausgewählt werden und dem Endnutzer kommuniziert (z.B. in den Metadaten des Datensatzes im Open-Data-Portal)

Hinweise

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

• Zeichencodierung

- Wenn ein Datensatz mit der “falschen” Codierung geöffnet wird, kommt es zu Probleme mit der Lesbarkeit des Datensatzes

- z.B.:

Adler, Alfred: Praxis und Theorie der Individualpsychologie. München			
Adler, Bruno: Sämtliche Schriften.			
Adler, Felix: Der Moralunterricht der Kinder. Berlin: Dümmler 1894.			
Adler, Friedrich (Wolfgang): Sämtliche Schriften.			
Adler, Georg: Sämtliche Schriften.			
Adler, Max: Sämtliche Schriften.			

Auswertung CSV

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- **Vorteil:** Am offensten
- **Nachteil:** Sehr einfache
Darstellungsmöglichkeiten

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Formatentscheidung: Excel vs. CSV

Wann sollte man was nutzen?

- In den meisten Fällen: CSVs
- Excel benutzen wenn...
 - es wichtig ist, dass die Tabellen lesbarer für Menschen sind (**aber denken Sie daran, dass die Tabelle maschinenlesbar bleiben muss!**)
 - mehrere Blättern nötig sind

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Andere Formate

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Daten Formate

- **Nicht tabellarisch**

- JSON
- GeoJSON

- **Tabellarisch**

- SAV
- DTA
- RData

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:

Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Daten Formate

• JSON

- “JavaScript Object Notation”
- Nicht tabellarisch, mehr flexibel

```
"Entity1" : {  
    "type" : "object",  
    "title" : "Entity1",  
    "properties" : {  
        "Attribute2" : {  
            "type" : "string",  
            "$ref" : "ReferencedDomains.json/properties/CharDomain"  
        },  
        "Attribute1" : {  
            "type" : "string",  
            "required" : true,  
            "maxLength" : 5,  
            "minLength" : 5  
        }  
    }  
}
```

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:

Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Daten Formate

• GeoJSON

- “Geographic JSON”
- Geografische Daten
- Basierend auf dem JSON-Format

```
1 {  
2   "type": "FeatureCollection",  
3   "features": [  
4     {  
5       "type": "Feature",  
6       "geometry": {  
7         "type": "Point",  
8         "coordinates": [-111.125, 33.375]  
9       },  
10      "properties": {  
11        "trackid": "AA-1234",  
12        "reported_dt": "12/31/2019 23:59:59"  
13      }  
14    },  
15    {  
16      "type": "Feature",  
17      "geometry": {  
18        "type": "Point",  
19        "coordinates": [-113.675, 35.875]  
20      },  
21      "properties": {  
22        "trackid": "AA-1234",  
23        "reported_dt": "12/31/2019 23:59:59"  
24      }  
25    }  
26  ]  
27 }
```

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Daten Formate

- SAV
 - SPSS Daten
- DTA
 - Python Daten
- RData
 - R Daten

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Häufige Probleme mit tabellarischen Datensätzen

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

5 Sachen zu überprüfen

① Zahlen

② Daten

③ Standorte

④ Tippfehler

⑤ Identifikationsnummer

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:

Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

1. Zahlen

- Für Ganzzahlen: Keine Leerzeichen oder Kommata verwenden



bewilligungssumme
248 400
200 000
4 731 300
288 116
293 766
341 343
6 376 500



bewilligungssumme
248400
200000
4731300
288116
293766
341343
6376500
374625

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

1. Zahlen

- **Für Dezimalzeichen:** Immer einen Punkt verwenden (0.1 vs. 0,1)



Haushaltsenergie (Strom, Gas und andere Brennstoffe)
99,8
99,9
99,7
99,9
99,1
98,4
98,4



Haushaltsenergie (Strom, Gas und andere Brennstoffe)
99.8
99.9
99.7
99.9
99.1
98.4
98.4

Mehr Dazu: NQDM: "Leitfaden für qualitativ hochwertige Daten und Metadaten", S. 24.

2. Daten

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- Empfehlung des NQDM:
ISO-8601-Format nutzen

- JJJJ-MM-TT (z.B.: 2020-09-17)



von	bis
05.09.20	05.09.20
19.09.20	20.09.20
09.08.20	09.08.20
22.08.20	22.08.20
01.08.20	01.08.20



von	bis
2020-09-05	2020-09-05
2020-09-19	2020-09-20
2020-08-09	2020-08-09
2020-08-22	2020-08-22
2020-08-01	2020-08-01

Mehr Dazu: NQDM: "Leitfaden für qualitativ hochwertige Daten und Metadaten", S. 20.

<https://www.nqdm-projekt.de/de/downloads/leitfaden>

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

3. Standorte

Zwei Möglichkeiten:

① Koordinaten

- Die beste Wahl (am genauesten)
- RBS-Geocoder (vom Amt für Statistik)

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

3. Standorte

② Adressen

- Empfehlung: Straße, Hausnummer und PLZ trennen

postleitzahl	strasse	hausnummer
10785	Potsdamer Straße	35
10117	Am Festungsgraben	1
10553	Rostocker Straße	32 b
10117	Am Kupfergraben	5
10178	Oranienburger Straße	30
10559	Lübecker Straße	13
10785	Schöneberger Ufer	57
13347	Reinickendorfer Straße	61
13347	Ruheplatzstraße	13

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:

Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

3. Standorte

② Adressen

- Darauf beachten: Adressen möglichst einheitlich erfassen
 - zB.: “Straße” vs. “Str.”

4. Tippfehler

- Mit der Anwendung eines Filters in Excel kann man Tippfehler leicht identifizieren
 - Données -> Filtre avancé -> Extraction sans doublon

Buenos-Aires
Buss/Saar
Cassarate
Cassarete
Celje
Celle

Quelle: Open Data Portal, "Liste der verbotenen Bücher"

- Rechtschreibkorrektur in Excel aktivieren

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

4. Tippfehler

- Technologiestiftung Tool zum Aufräumen eines CSV

Wert	Anzahl	Ersetzen durch
Berlin	3	<input type="radio"/>
Brlin	2	<input type="radio"/>
'Berlin'	1	<input type="radio"/>
Börlen	1	<input checked="" type="radio"/>
Brln	1	<input type="radio"/>
Hamburg	1	<input type="radio"/>
Hamburh	1	<input checked="" type="radio"/>

<https://lab.technologiestiftung-berlin.de/projects/csv-string-optimization/de/>

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

5. ID Nummer

- Für Datensätze, wo die Angaben eine Bezeichnung/ID-Nummer haben: diese Nummer sollte im Datensatz drin sein!
 - z.B. Die Bezeichnungsnummer einer Schule
 - Wichtig um...
 - Änderungen zu verfolgen
 - Datensätze zusammenzuführen

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Das wichtigste von allen: Einheitlichkeit

- Wählen Sie einen Standard aus - am besten einen von uns empfohlenen Standard - und **ziehen Sie diesen Standard konsequent durch**

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

Quiz

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

▶ Datensätze

Datensätze	Schulbausanierung Sommer 2020
Arbeitsmarkt	Berlin investiert in den kommenden Jahren 5,5 Milliarden Euro in die Sanierung und in den Bau von Schulgebäuden. Auch in den Sommerferien 2020 gehen an den Berliner Schulen die Bau- und Sanierungsmaßnahmen weiter.
Bildung	
Demographie	
Geographie und Stadtplanung	
Gesundheit	
Jugend	
Kunst und Kultur	
Öffentliche Verwaltung, Haushalt und Steuern	
Protokolle und Beschlüsse	
Sonstiges	

Lizenz: Creative Commons Namensnennung 

Kategorie: Bildung

Geographische Abdeckung: Berlin

Geographische Granularität: Berlin

Zeitliche Granularität: Keine

Veröffentlicht: 27.07.2020

Aktualisiert: 27.07.2020

Veröffentlichende Stelle: Senatsverwaltung für Bildung, Jugend und Familie

E-Mail Kontakt: karen.koenig AT senbjf.berlin.de

Website: <https://www.berlin.de/sen/bildung/service/daten/>

Tags: Bauen | Sanieren | Schulbaumaßnahmen ... (2 weitere)

Kommentare: 0

Fragen

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:

Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- In Excel, sortieren Sie die Eingaben nach Finanzierung. Merken Sie etwas, was verbessert sein könnte?
- Schauen Sie sich die Adressen in der “Adresse” Spalte an. Was könnte hier optimiert werden?
- Wie kann man den Datensatz zu einem CSV konvertieren und speichern?

Analyse

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:

Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

• Finanzierung-Spalte

- Einheitlichkeit ist wichtig
- Auf Typos achten!

Kosten in Euro	Finanzierung
1.175.000	1.254.000
1.262.000,00	baul. Unterhaltung
415.000,00	BENE, Baubudget geplant
1.920.000,00	Eigener Bezirkshaushalt
31.000.000,00	Eigener Bezirkshaushalt
1.066.000	102.000
3.775.000,00	Eigener Bezirkshaushalt
190.215	809.831 962.553
519.929	244.119
648.748	11.652 249.908
98.250	249.908
23.712	475.277
720.000,00	Investitionsprogramm
1.000.000,00	Investitionsprogramm
1.500.000	800.000
2.469.867,00	Investitionsprogramm (Ausweichkosten)
280.000,00	Investitionsprogramm SenStadtWohn
800.000	1.028.000
110.000,00	Investitionsprogramm
900.000	1.100.000
500.000	1.100.000
800.000,00	Investitionsprogramm (Ausweichkosten)
430.000,00	SchulISP
1.500.000,00	Eigener Bezirkshaushalt
599.000,00	Kommunalinvestitionsprogramm
10.000.000	265.000
200.000,00	Stadtbaulicher Datumsentschädigung
330.000	200.000
330.000	SchulISP, Eigener Bezirkshaushalt
1.200.000,00	BWA
950.000,00	Kommunalinvestitionsprogramm
1.942.681,00	SchulISP
415.000,00	Eigener Bezirkshaushalt
2.800.000,00	Geldanlagen im Gewerbeabschutz
136.787	Baubudget geplant
27.329	Baubudget geplant
564.317	PV LAGEBO, Baubudget geplant
400.000,00	Eigener Bezirkshaushalt
775.000,00	Eigener Bezirkshaushalt
400.000,00	Eigener Bezirkshaushalt

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

• Adresse-Spalte

- “Straße” oder “Str.”?

Adresse
Ungarnstraße 75
Guineastrasse 17-18
Lützowstraße 83-85
Müllerstraße 158
Berolinastraße 8
Schöningstraße 17
Reinickendorfer Straße 60
Alt-Moabit 10
Bochumer Str. 8
Osloer Str. 23-26
Schwyzer Str. 6-8

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

3. Beispiel

Gieß den Kiez

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel



Gieß den Kiez

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

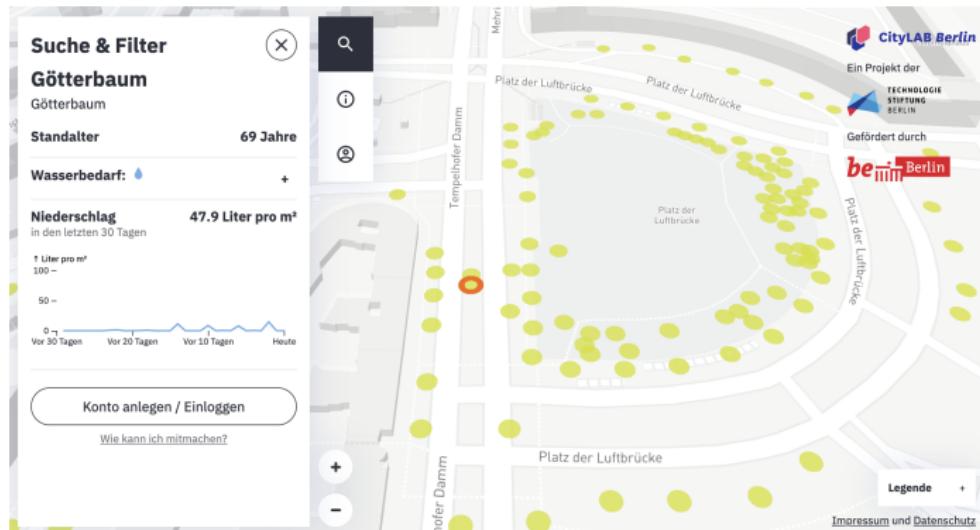
Formatentscheidung:
Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel



Fehlende ID Nummer

Einführung

1. Theorie

Quiz

2. Praxis

Excel

CSV

Formatentscheidung:

Excel vs. CSV

Andere Formate

Häufige Probleme
mit tabellarischen
Datensätzen

Quiz

3. Beispiel

- **Ziel:**

- Bewässerungsdaten von der Regierung in den Gieß den Kiez Datensatz hinzufügen

- **Problem:**

- Kein gemeinsame Identifikationsnummer zwischen:

① Bewässerungdaten von der Regierung

② Bewässerungdaten von Gieß den Kiez

Versuch #1

- **Methode:** ID aus verschiedenen Spalten generieren
 - Pflanzjahr + Standalter + Stammumfang + Stammdurchmesser + Baumhöhe + Kronendurchmesser + Sorte

```
bewaesserung$id <- NA  
bewaesserung$id <- paste0(bewaesserung$Pflanzjahr,  
                           bewaesserung$Standalter,  
                           bewaesserung$Stammumfang,  
                           bewaesserung$Stammdurchmesser,  
                           bewaesserung$Baumhöhe,  
                           bewaesserung$Kronendurchmesser,  
                           bewaesserung$Sorte.deutsch)
```

- **Problem:** Duplikate (i.e. Bäume mit identischen Charakteristiken im Datensatz)