CORRELAID

GOOD CAUSES. BETTER EFFECTS.

Grundlagen Datenmanagement

Praktische Übung: Alles Excel?

- 60 min -

Für kleine Datensätze, die nur selten Updates benötigen, ist oft auch Excel ausreichend

Excel

Excel bietet als kostengünstiges Office-Tool eine gute Möglichkeit kleine Datensätze, an denen nur eine oder sehr wenige Personen arbeiten, abzubilden.

\+
- Onlineversionen ermöglichen Kollaborationen
- Geringe Zugangshürden für versch. Nutzer:innen
- Integration mit fast allen Tools
- Kostengünstige Anschaffung

−
- Rudimentäre Eingabekontrolle
- Schutz- und Nutzungsrechte oft nicht ausreichend
- Geringes Automatisierungspotenzial für Reports
- Funktionalitäten begrenzt

Wie können wir hier das volle Potenzial des Tools ausschöpfen?

Datenbankmanagementsystem

Bei Datensammlungen, die organisationsweit genutzt werden und regelmäßig angepasst werden, lohnt sich die Anschaffung einer Datenbanklösung.

\+
- I.d.R. fortgeschrittene Schutz- und Nutzungsrechte
- Komplexe Abfragen möglich
- Automatisierung von Reportprozessen
- Verschiedene Datenansichten
- Umfangreiche Eingabekontrolle

−
- Schnittstellen zur Toolintegration notwendig
- Umfangreiche Planung nötig
- Hoher Kostenfaktor
- Verwaltung- und Maintenance-Anforderungen hoch

Excel - wer kennt es nicht?

34 Euro kostet eine Office Standard Lizenz über Stifter-helfen (zzgl. MwSt.)

85 Prozent der Unternehmen (n = 1.023) in Deutschland nutzten 2020 MS Office

88 Prozent der in 13 Audits untersuchten Excel-Arbeitsmappen enthielten Fehler

80 Prozent der Fehler können in Qualitätsprüfungen in Zelle-bei-Zelle-Inspektionen eliminiert werden (!)

Trotz allen Herausforderungen: Excel bleibt wohl erst so prominent, wie es ist. Deshalb sollten wir lernen, wie man gut damit arbeitet und übliche Fehler vermeidet.

Quellen: Meistgenutzte Office-Software in Unternehmen in Deutschland 2020, Statista Research Department, Nov. 2020, abgerufen am 04.05.21 unter diesem Link.
Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: 10.1080/00031305.2017.1375989.

Acht Prinzipien schlug Tom Grossmann bereits 2002 für das Aufsetzen von Arbeitsmappen vor

1 **Standards folgen**
Empfohlenen Vorgehensweisen zu folgen, hat eine große Wirkung.

2 **Planen**
Die Planung des Lebenszyklus einer Arbeitsmappe ist wichtig.

3 **Vorbereiten**
A priori (im Vorhinein) definierte Anforderungen sind förderlich.

4 **Vorausschauen**
(Auch) die zukünftige Nutzung der Excel-Arbeitsmappe ist wichtig.

5 **Designen**
Das Design sollte ebenfalls durchdacht werden (Vorsicht: Form folgt Funktion!).

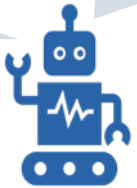
6 **Flexibel sein**
Die empfohlenen Vorgehensweisen sind situationsabhängig und sollten flexibel sein.

7 **Kollaborieren**
Die Programmierung von Arbeitsmappen sollte in Teams erfolgen.

8 **Kosten einplanen**
Die Entwicklung von empfohlenen Vorgehensweisen ist schwer/verbraucht Ressourcen.

Ihren Ursprung haben diese Prinzipien im Software Engineering.

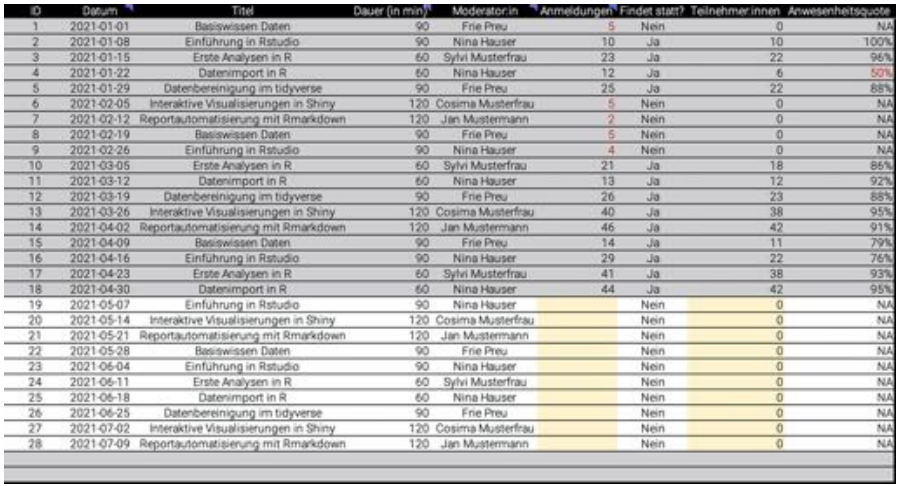
Weitere Referenzen der EuSpRIG findet Ihr hier!



Quelle: Thomas A. Grossman, Spreadsheet Engineering: A Research Framework, European Spreadsheet Risks Interest Group Symposium, Wales, July 2002, abgerufen am 04.05.21 unter diesem Link.

CORRELAID
GOOD CAUSES. BETTER EFFECTS.

Eine gutes Excel-Arbeitsblatt hat eine ähnliche Struktur wie die Daten in einer Datenbank…



ID	Datum	Titel	Dauer (in min)	Moderator:in	Anmeldungen	Findet statt?	Teilnehmer:innen	Anwesenheitsquote
1	2021-01-01	Basiswissen Daten	90	Frie Preu	5	Nein	0	NA
2	2021-01-08	Einführung in Rstudio	90	Nina Hauser	10	Ja	10	100%
3	2021-01-15	Erste Analysen in R	60	Sylvi Musterfrau	23	Ja	22	96%
4	2021-01-22	Datenimport in R	60	Nina Hauser	12	Ja	6	50%
5	2021-01-29	Datenbereinigung im tidyverse	90	Frie Preu	25	Ja	22	88%
6	2021-02-05	Interaktive Visualisierungen in Shiny	120	Cosima Musterfrau	5	Nein	0	NA
7	2021-02-12	Reportautomatisierung mit Rmarkdown	120	Jan Mustermann	2	Nein	0	NA
8	2021-02-19	Basiswissen Daten	90	Frie Preu	5	Nein	0	NA
9	2021-02-26	Einführung in Rstudio	90	Nina Hauser	4	Nein	0	NA
10	2021-03-05	Erste Analysen in R	60	Sylvi Musterfrau	21	Ja	18	86%
11	2021-03-12	Datenimport in R	60	Nina Hauser	13	Ja	12	92%
12	2021-03-19	Datenbereinigung im tidyverse	90	Frie Preu	26	Ja	23	88%
13	2021-03-26	Interaktive Visualisierungen in Shiny	120	Cosima Musterfrau	40	Ja	38	95%
14	2021-04-02	Reportautomatisierung mit Rmarkdown	120	Jan Mustermann	46	Ja	42	91%
15	2021-04-09	Basiswissen Daten	90	Frie Preu	14	Ja	11	79%
16	2021-04-16	Einführung in Rstudio	90	Nina Hauser	29	Ja	22	76%
17	2021-04-23	Erste Analysen in R	60	Sylvi Musterfrau	41	Ja	38	93%
18	2021-04-30	Datenimport in R	60	Nina Hauser	44	Ja	42	95%
19	2021-05-07	Einführung in Rstudio	90	Nina Hauser		Nein	0	NA
20	2021-05-14	Interaktive Visualisierungen in Shiny	120	Cosima Musterfrau		Nein	0	NA
21	2021-05-21	Reportautomatisierung mit Rmarkdown	120	Jan Mustermann		Nein	0	NA
22	2021-05-28	Basiswissen Daten	90	Frie Preu		Nein	0	NA
23	2021-06-04	Einführung in Rstudio	90	Nina Hauser		Nein	0	NA
24	2021-06-11	Erste Analysen in R	60	Sylvi Musterfrau		Nein	0	NA
25	2021-06-18	Datenimport in R	60	Nina Hauser		Nein	0	NA
26	2021-06-25	Datenbereinigung im tidyverse	90	Frie Preu		Nein	0	NA
27	2021-07-02	Interaktive Visualisierungen in Shiny	120	Cosima Musterfrau		Nein	0	NA
28	2021-07-09	Reportautomatisierung mit Rmarkdown	120	Jan Mustermann		Nein	0	NA

1) **Variablen** und was eine **Beobachtung** ist wird von Anfang an **definiert**.
2) Die **Variablen** werden in der **waagerechten** erweitert. Insgesamt bleiben sie **übersichtlich**.
3) Die **Beobachtungen** werden in der **senkrechten** hinzugefügt.

…und folgt ähnlichen Grundkonzepten

Konsistenz

Die **Dateneingabe und –Organisation** sollte **konsistent** erfolgen. Das gilt insb. für: Die Codierung von kategorischen Variablen und fehlenden Werten, Variablenbezeichnungen, IDs, Datenformaten, Layouts, Dokumentenbezeichnungen und Notizen. Achte auf **nicht-notwendige Leerzeichen**. **Sperre** finale Zellen.

Klarheit

Bei der **Auswahl von Codierungen und der Benennung von Variablen** sollte darauf geachtet werden, dass diese **klar und kurz** sind. Einige Datenwissenschaftler:innen bevorzugen die Trennung von mehreren Wörtern in Bezeichnungen mit Unterstrichen „_“ (und ohne Leerzeichen und spezielle Charaktere).

Standards

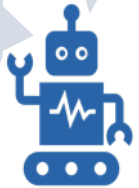
Nutze **die Datenvalidierung, Formatvorgaben und Formeln**. Für Datumsformate ist der Standard ISO 8601 (**YYYY-MM-DD**) zu empfehlen, da dieser in den meisten Programmiersprachen genutzt wird. Lasse Zellen nicht leer, sondern nutze „**NA**“. Kreiere ein **Datenverzeichnis**, indem Datenformate und Variablenbedeutung stehen.

Form

Eine Zelle sollte nur **eine Dateninstanz** beinhalten. **Einheiten** gehören **in** den **Variablennamen**. Diese sollten alle in einer Zeile stehen (**keine zweizeiligen Header**!). Die **Gesamtform** des Arbeitsblatts sollte **rechteckig** sein. Nutze pro Datensatz **einen Reiter** und einen **separaten für Analyse**. **Formatierung** hat **nur in Excel** Bedeutung.

Andere Software liest nur aus, was explizit in der Excel steht. Als Test kannst du Excel in CSV konvertieren und schauen, ob die enthaltenen Informationen noch dieselben sind.

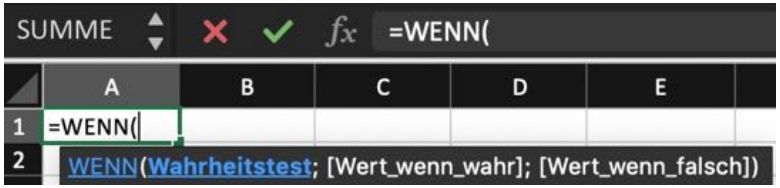
Hast du schon an das Back-up gedacht? Willst du Version Control erlauben?



CORRELAID
GOOD CAUSES. BETTER EFFECTS.

Quelle: Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: 10.1080/00031305.2017.1375989, abgerufen am 04.05.21 unter diesem Link.

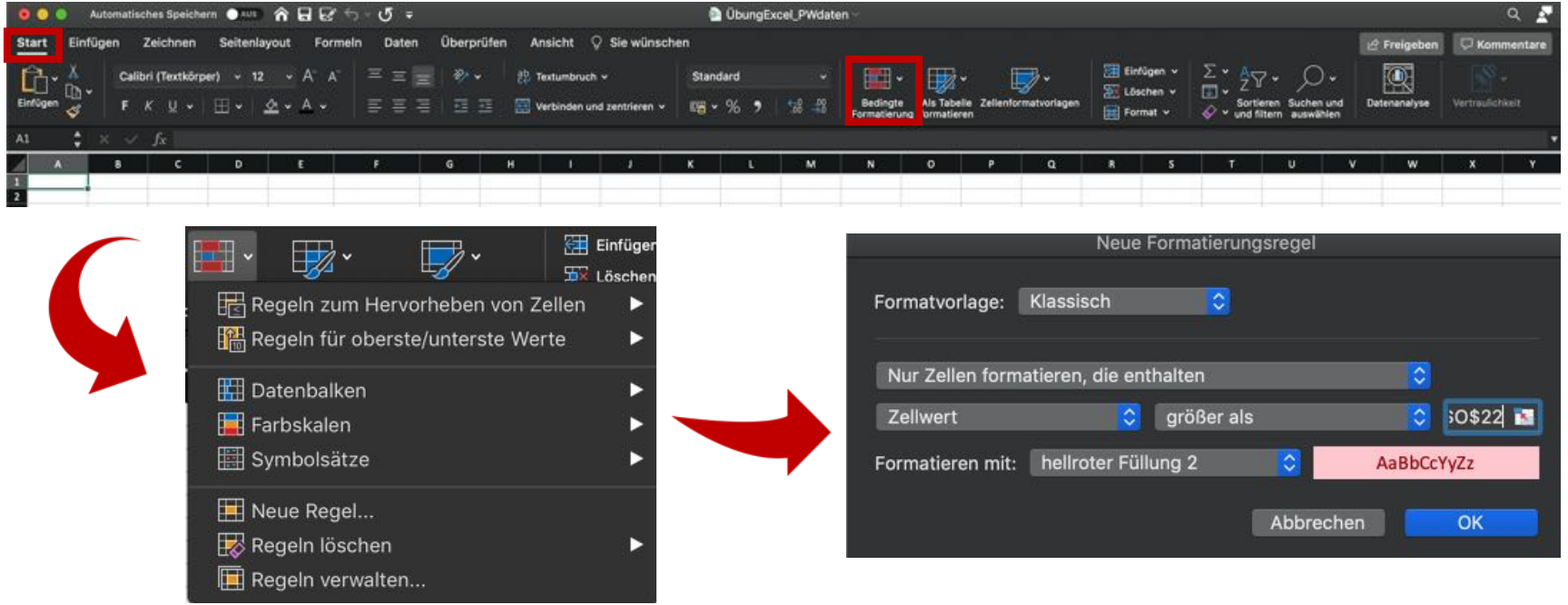
Was mit Formeln berechnet oder ermittelt werden kann, sollte so berechnet oder ermittelt werden



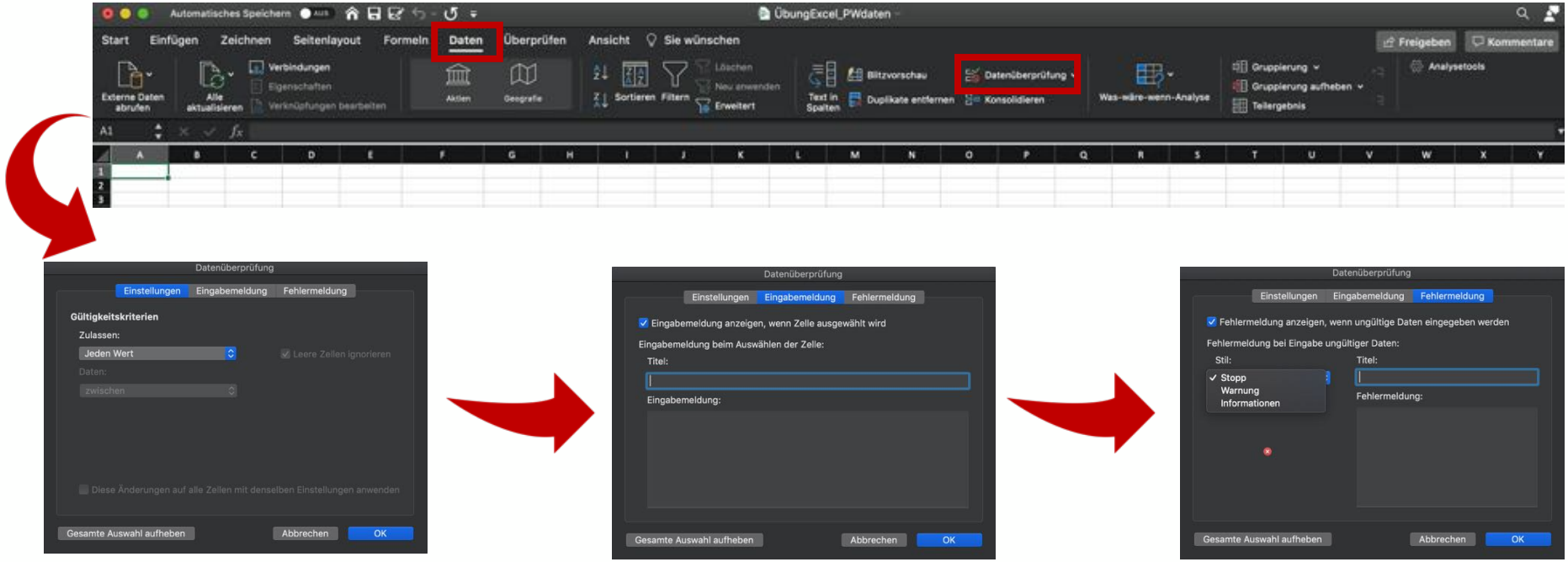
Wie die Funktionen angewendet werden, findet Ihr hier!

CORRELAID
GOOD CAUSES. BETTER EFFECTS.

Die leichte Form der Eingabekontrolle ist die bedingte Formatierung

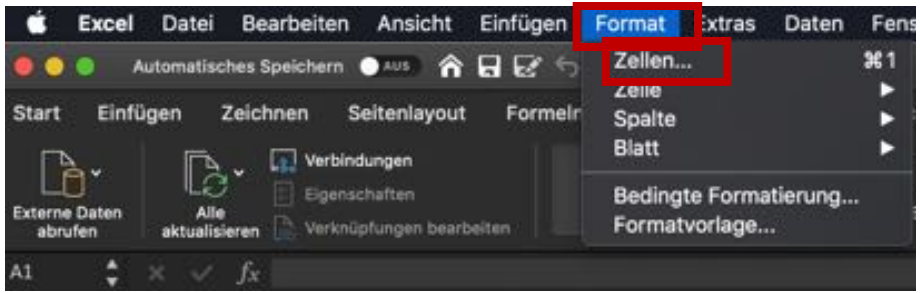


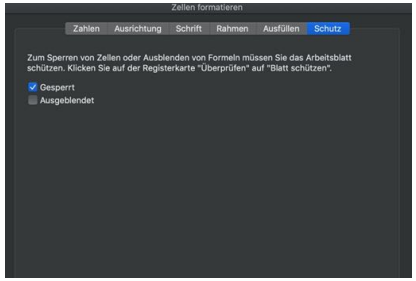
Die harte Form der Eingabekontrolle ist die Datenvalidierung



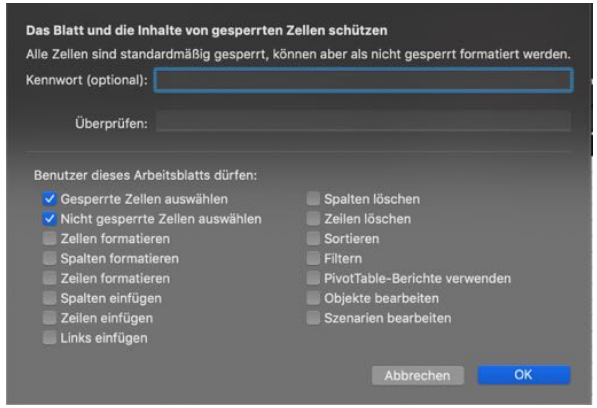
Damit Daten nicht später verändert oder gelöscht werden, ist es sinnvoll Zellen zu sperren

1

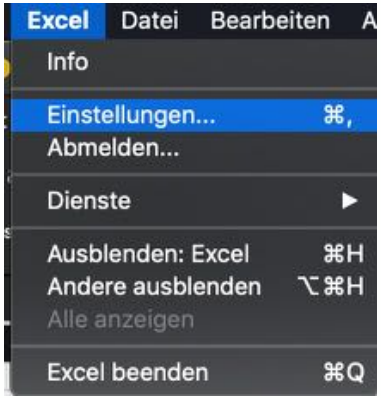


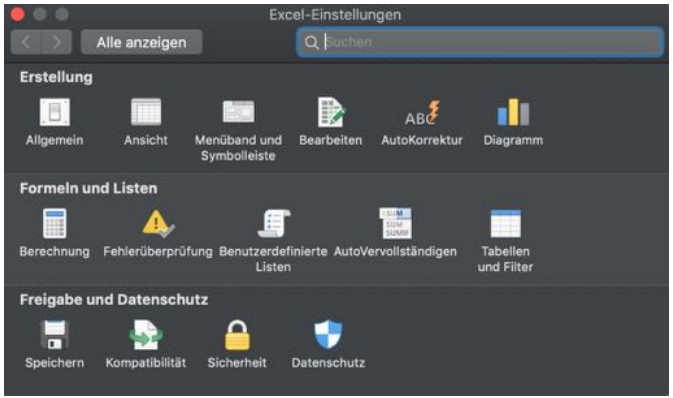


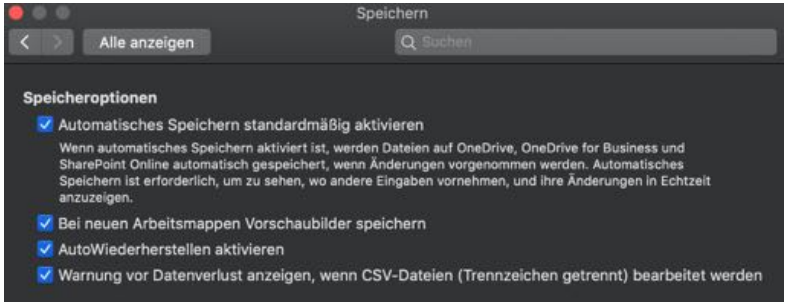
2



Eine Form der Versionskontrolle gibt über Speichereinstellung und die Nutzung von OneDrive





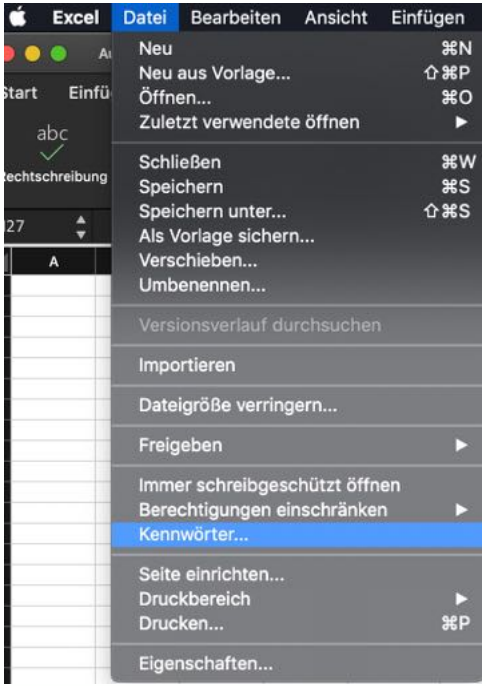


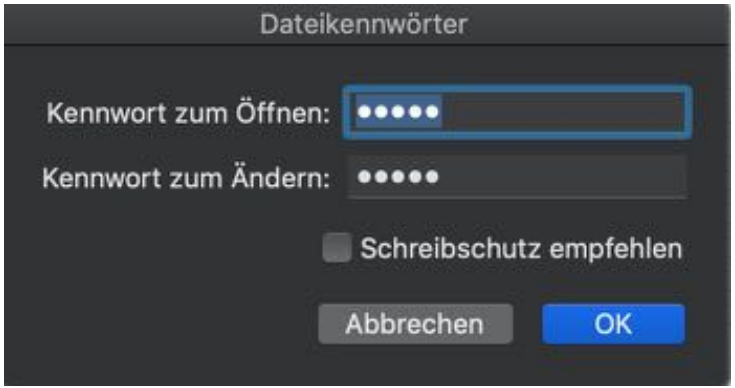
OneDrive

Problem: Datenschutz

Das Risiko ist allerdings bei OneDrive-Servern in der EU gering und kann mit einer Verschlüsselung fast vollständig eliminiert werden. Auch wir haben dazu die Berliner Datenschutzbeauftragen befragt. Mehr Informationen dazu unter EuGH-Urteil Schrems II. Die Anleitung zur Einrichtung gibt es hier.

Aus Datenschutzgründen sollten Excel mit sensiblen Daten passwortgeschützt sein





Ob das ausreicht, um das Schrems II Urteil zu umgehen, ist zweifelhaft.

Und so kann eine Excel dann aussehen: