Einführung in Apache Superset: Ein Chart

Ein Arbeitsblatt für Interessierte und Lehrpersonen

Übersicht

Lernziele

Diese Einführung hat folgende Lernziele:

- Du kennst die wichtigsten Konzepte und Grundfunktionen von Apache Superset
- Du kannst Datenquellen selektieren
- Du kannst Daten mittels Charts visualisieren
- Du kannst Charts zu einem Dashboard anordnen
- Du kannst Dashboards mit anderen (übers Web) teilen

Das Bearbeiten dieses Arbeitsblatts dauert ca. eine halbe Stunde, je nach deinem Vorwissen.

Für diese Einführung werden keine Programmierkenntnisse vorausgesetzt; Grundkenntnisse der Tabellenkalkulation genügen fürs Erste.

Für die Übungen in diesem Arbeitsblatt benötigst du Zugang zu einem Apache Superset-Service (vgl. unten), sowie einen gängigen Webbrowser (am besten funktioniert Superset auf Chrome) und eine Internetverbindung.

Diese Anleitung bezieht sich auf Release 2.0 von Apache Superset.

Einleitung

Für den Erfolg einer Entscheidungsfindung ist eine gute Visualisierung wichtig. Darum müssen die Erkenntnisse eines Anliegens, Projektes oder Umfrage visualisiert werden. Wenn Erkenntnisse anschaulich und nachvollziehbar dargestellt werden, erhöht das deren Verständlichkeit und Akzeptanz.

Visualisierung kann man zudem nicht nur zur Veranschaulichung einsetzen, sondern auch zur Datenanalyse. Häufig erkennt man Zusammenhänge in den Daten erst durch eine geschickte Darstellung. Wir Menschen sind schlecht darin, Zahlen zu vergleichen; grafische Muster dagegen erkennen wir gut. Visualisierung stellt somit nicht nur die Daten grafisch dar, sondern kann auch als eigene Technik der Datenanalyse eingesetzt werden. Das Internet ermöglicht zudem die Publikation der Visualisierung und damit die einfache Kommunikation mit Kunden und Arbeitskollegen.

Apache Superset ist so ein Daten-Visualisierungs- und Publikations-Werkzeug. Manche nennen diese Anwendung auch "Business Intelligence Tool". Superset ist auch ein Werkzeug zum Teilen (Sharing) von Datenquellen, d.h. von Tabellen-Daten bis zu Geodaten. Es kann mit verschiedensten Datenbanken verbunden werden.

Zu Apache Superset gibt es eine offizielle, englischsprachige Dokumentation.

Das FAQ (Frequently Asked Questions, Deutsch: "häufig gestellte Fragen"), welches auf der Seite zu finden ist, enthält teils nützliche Informationen.

Erstellen einer Apache-Superset-Instanz

Für die folgenden Übungen wird mithilfe des coders eine Server-Instanz von Apache-Superset verwendet. Dabei muss man sich zuerst mit seinem OST GitLab Konto anmelden. Schliesslich muss unter **Workspaces › + Create Workspace...** das “Superset” Template gewählt werden und dem Workspace einen geeigneten Namen gegeben werden.

Mit einem click **[Create Workspace]** wird nun eine Instanz von Apache-Superset erstellt und gestartet.

Alternativ kann auch lokal eine Apache-Superset-Instanz mit Docker aufgesetzt werden. Hierzu gibt es ein Tutorial auf der offiziellen Superset-Website.

Konzepte und Begriffe

Nach dem Anmelden an einer Apache Superset-Instanz (siehe vorhergehendes Kapitel), wird die Startseite von Apache-Superset angezeigt.

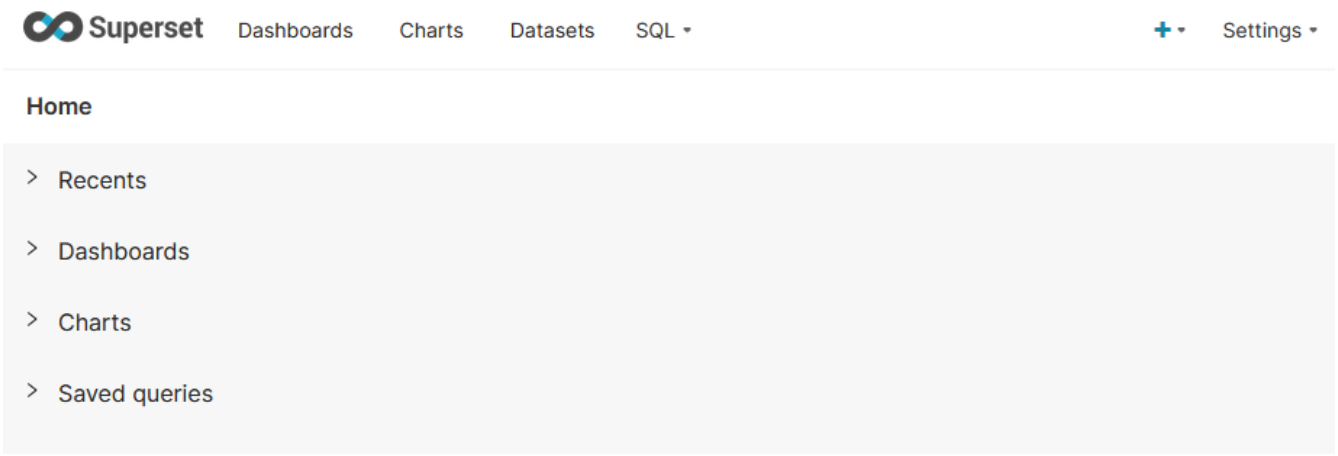

Abbildung 1. Startbildschirm von Apache-Superset.

Hier einige Erläuterungen zu den Konzepten hinter Apache Superset:

Data (Datenquelle)

Dies sind Datenbanken, welche anschliessend ausgewertet und verarbeitet werden. Auf diese können später dann auch SQL-Abfragen getätigt werden.

Chart (Diagramm)

In einem *Chart* werden die Daten aus einem Dataset grafisch dargestellt.

Dashboard

Eine Zusammenstellung aus verschiedenen Charts, welche als Website aufgerufen werden kann.

Metric

Eine "Metric", manchmal auch ein "Measurement" genannt, ist eine numerische Kennzahl. Metrics / Measurements werden v.a. in den Charts erwähnt und verlangt.

Record

Datenquellen und Charts sind alles Programmierelemente, die manchmal in der Benutzeroberfläche als "Record" bezeichnet werden.

SQL-Query

Anweisung, Datenbankanfrage in der Datenbanksprache SQL. SQL-Queries kann man speichern und als Datenquelle anderen zur Verfügung stellen.

Daten und Fragestellungen

Die mit Apache Superset (und Business Intelligence Tools allgemein) zu visualisierenden Daten müssen in strukturierter und sauberer Form vorliegen. Wenn nötig, müssen die Daten mit Datenbanksystemen (SQL), Tabellenkalkulationsprogrammen (z.B. MS Excel, LibreOffice) oder GIS (z.B. QGIS) aufbereitet werden (siehe u.a. OpenSchoolMaps > "Einführung in QGIS 3 und in Geoinformationssysteme (GIS)").

Für die Bereinigung von Daten ("Data Wrangling") gibt es OpenRefine, welches ein Programm ist, welches den Prozess erheblich vereinfacht. Hierzu gibt es ebenfalls ein Arbeitsblatt auf OpenSchoolMaps > "Daten sichten, bereinigen und integrieren mit OpenRefine".

In diesem Arbeitsblatt wird nur eine Tabelle verwendet, die Tabelle `wb_health_population` von der Weltbank (Quelle, Lizenz CC BY-4.0, Stand ca. 2017, übersetzt etwa "Weltbank-Gesundheit-Bevölkerung").

Die Tabelle `wb_health_population` hat ca. 328 Spalten (Attribute), d.h. sehr viele. Wir verwenden davon folgende Spalten:

- `country_name`: Name des Landes
- `region`: Weltregion, in der das Land liegt
- `year`: Jahr der Datenerhebung (1960 - 2014)
- `SP_POP_TOTL`: Anzahl Einwohner insgesamt

Abbildung 2 zeigt die Daten, die wir nutzen werden. Nimm dir doch kurz Zeit und schaue dir diese Daten genau an. Zu verstehen, welche Daten in welcher Spalte sind, ist eine Notwendigkeit, um sinnvolle und korrekte Diagramme erstellen zu können.

region	country_name	year	SP_POP_TOTL
Europe & Central Asia	Switzerland	1960-01-01T00:00:00	5327827
Europe & Central Asia	Switzerland	1961-01-01T00:00:00	5434294
Europe & Central Asia	Switzerland	1962-01-01T00:00:00	5573815
Europe & Central Asia	Switzerland	1963-01-01T00:00:00	5694247
Europe & Central Asia	Switzerland	1964-01-01T00:00:00	5789228
Europe & Central Asia	Switzerland	1965-01-01T00:00:00	5856472
Europe & Central Asia	Switzerland	1966-01-01T00:00:00	5918002
Europe & Central Asia	Switzerland	1967-01-01T00:00:00	5991785
Europe & Central Asia	Switzerland	1968-01-01T00:00:00	6067714
Europe & Central Asia	Switzerland	1969-01-01T00:00:00	6136387
Europe & Central Asia	Switzerland	1970-01-01T00:00:00	6180877
Europe & Central Asia	Switzerland	1971-01-01T00:00:00	6213399
Europe & Central Asia	Switzerland	1972-01-01T00:00:00	6260956
Europe & Central Asia	Switzerland	1973-01-01T00:00:00	6307347
Europe & Central Asia	Switzerland	1974-01-01T00:00:00	6341405
Europe & Central Asia	Switzerland	1975-01-01T00:00:00	6338632
Europe & Central Asia	Switzerland	1976-01-01T00:00:00	6302504
Europe & Central Asia	Switzerland	1977-01-01T00:00:00	6281174
Europe & Central Asia	Switzerland	1978-01-01T00:00:00	6281738

Abbildung 2. Daten der Schweiz von 1960 - 1978 aus der Tabelle `wb_health_population`.

Charts (Diagramme)

Apache Superset bietet eine Vielzahl an verschiedenen Diagrammen. Im Anhang findest du weitere Beispieldiagramme zur Tabelle `wb_health_population`, die in weiterführenden Arbeitsblättern vorgestellt werden (Siehe Kapitel "Abschluss").

Ziel dieser Aufgabe ist es, ein Dashboard zu erstellen, das man mithilfe des Filters beeinflussen kann.

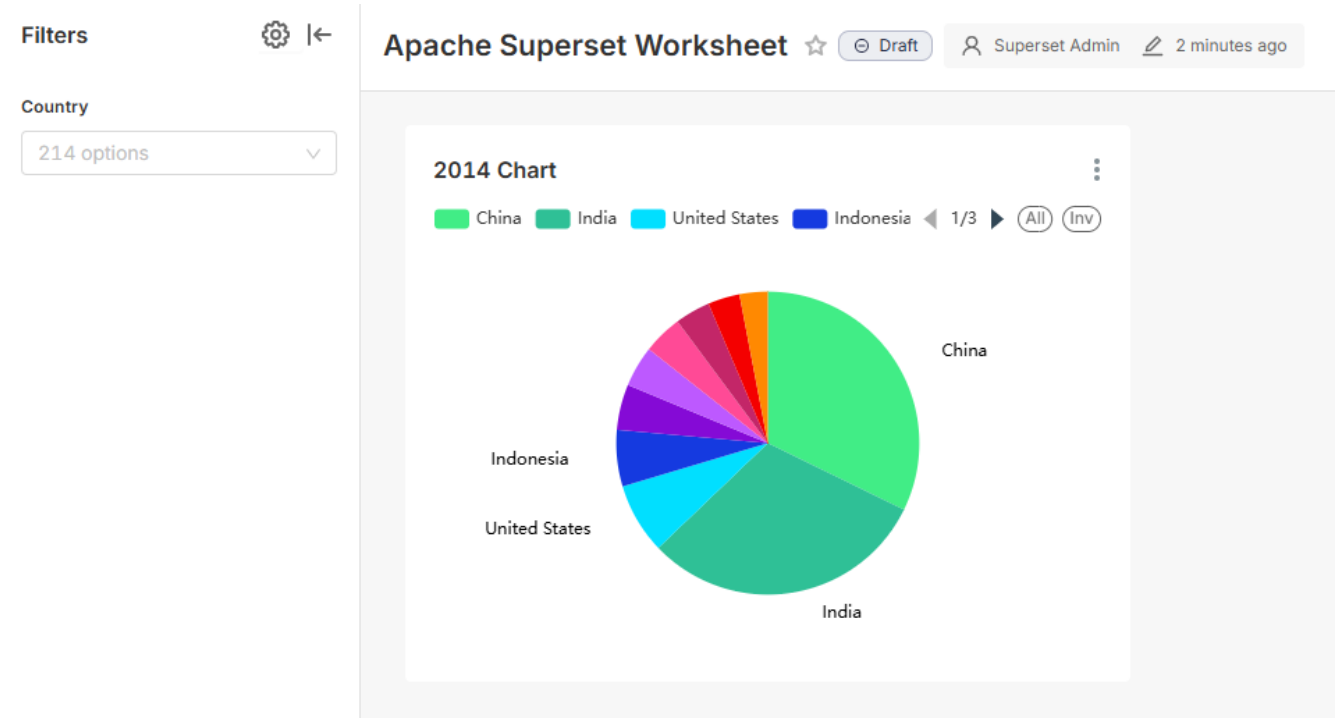

Abbildung 3. Das Dashboard mit dem Kreis-Diagramm und dem Filter.

Wenn du Änderungen an deinem Diagramm vornimmst, kannst du diese mit einem Klick auf **[Update Chart]** aktualisieren.

Aufgabe 1: Mein erster Chart

Um ein neues Diagramm zu erstellen, klicke auf **Charts › + Chart**.

Abbildung 4. Knopf zum Erstellen eines neuen Charts.

In dem neuen Fenster muss nun zunächst das gewünschte *Dataset* gewählt werden. Im Fall dieser Übung ist das `wb_health_population`.

Man kann jeweils nur eine Tabelle als Quelle auszuwählen. Möchte man mehrere Tabellen zu einer Datenquelle verknüpfen, benötigt man SQL-Kenntnisse. In weiteren Arbeitsblättern auf OpenSchoolMaps wird gezeigt, wie das geht.

Sobald ein Datenset ausgewählt wurde, kann nun in einem zweiten Schritt der gewünschte Diagrammtyp gewählt werden. Für diese Aufgabe wird ein `Pie Chart` (Kreis- oder auch Kuchendiagramm) benötigt.

Die Auswahl kann mit einem Klick auf **[CREATE NEW CHART]** bestätigt werden

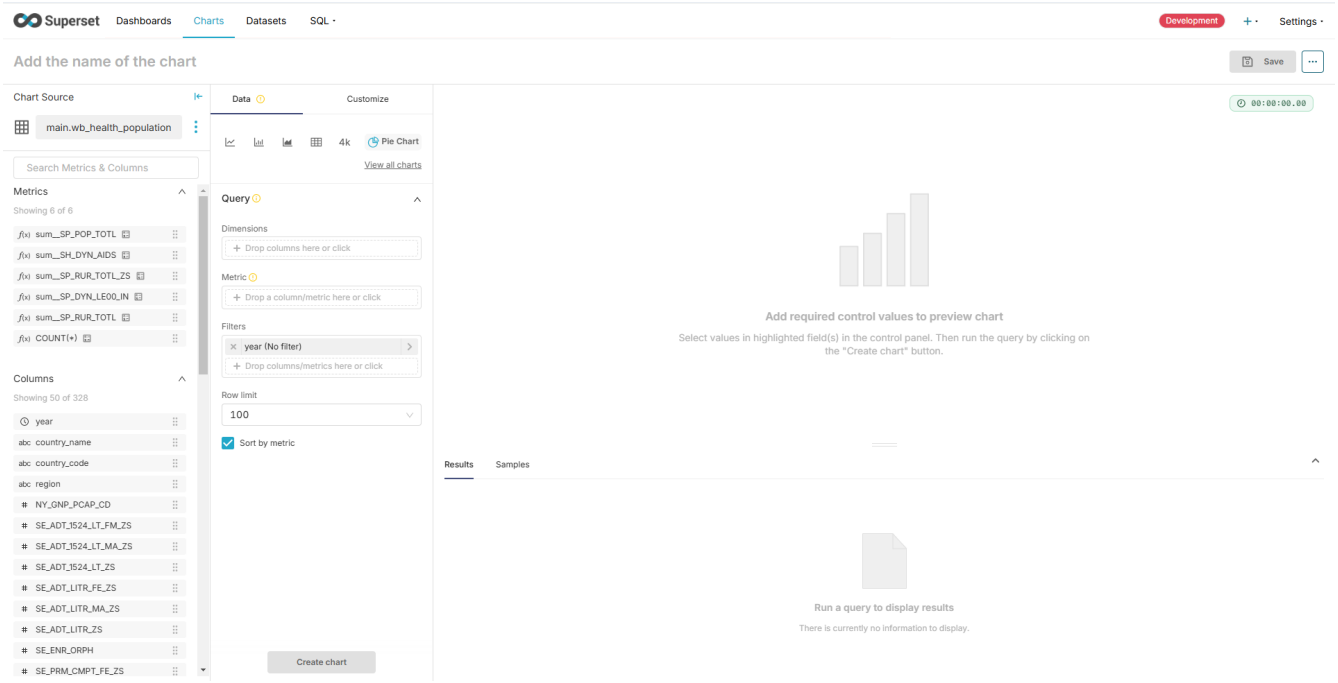

Abbildung 5. Oberfläche zur Konfiguration eines Charts.

Als erstes muss ausgewählt werden auf welche Werte das Diagramm aufgeteilt wird. (In welche Kuchenstücke das Diagramm aufgeteilt werden soll) Dies kann unter dem Punkt **Query › Dimensions** definiert werden. In diesem Beispiel sollen die einzelnen Abschnitte die Länder repräsentieren. Daher sollte hier `country_name` eingetragen werden.

Das Diagramm kann jedoch erst erstellt werden, wenn der Wert **Query › Metric** definiert wurde. Die `Metric` definiert den Wert, an welchem die Grösse der einzelnen Felder berechnet wird. In unserem Fall wollen wir die Bevölkerungsgrösse der Länder herausfinden, darum wird hier “MAX(SP_POP_TOTL)” eingefüllt.

Um diese Auszuwählen, klicke auf **METRIC › SIMPLE › COLUMN** und wähle `#SP_POP_TOTL`. Zusätzlich zu der Spalte muss noch der Wert **AGGREGATE** gesetzt werden. In unserem Fall soll `MAX` gewählt werden.

Nun können die Tabellen-Daten noch (zeitlich) gefiltert werden. Hierzu kann zunächst im Bereich **Filters › year (No filter)** der Bereich abgesteckt werden, welcher im Diagramm angezeigt werden soll.

Dabei sollte die **Time Range › RANGE TYPE** von `No filter` auf `Custom` umgestellt werden.

Nun kann die Zeit umgestellt werden. Hierbei kann zwischen verschiedenen Optionen gewählt werden. Es können entweder relative Zeitabstände (`Relative Date/Time`), absolute Zeitabstände (`Specific Date/Time`), bis zum Ende des Tages (`midnight`) oder bis jetzt (`now`) gewählt werden.

Um alle Daten zu erhalten - egal zu welchem Jahr sie gehören - oder wenn es keine Daten gibt, die zeitabhängig sind, kann man `No filter` wählen.

Wenn ein absoluter Zeitraum ausgewählt wird, muss die Eingabe mit einem Klick auf **[OK]** bestätigt werden. Ansonsten wird der Wert nicht übernommen.

Unter **Row Limit** kannst du das Ergebnis auf eine bestimmte Anzahl Einträge beschränken. Wenn du z.B. ein *Row Limit* von 10 setzt, werden nur die 10 bevölkerungsreichsten Länder angezeigt.

Wenn du jetzt **[CREATE CHART]** drückst, wird die Abfrage ausgeführt. Sie zeigt in einem Kreis-Diagramm die zehn bevölkerungsreichsten Länder.

Nun kann oben im Bereich **Add the name of the chart** ein Name für das Chart gewählt werden und das Diagramm mit **[SAVE]** (oben rechts) gespeichert werden.

Charts zu einem Dashboard anordnen

Im Reiter Dashboard kann unter **Dashboards › + DASHBOARD** ein neues Dashboard erstellt werden.

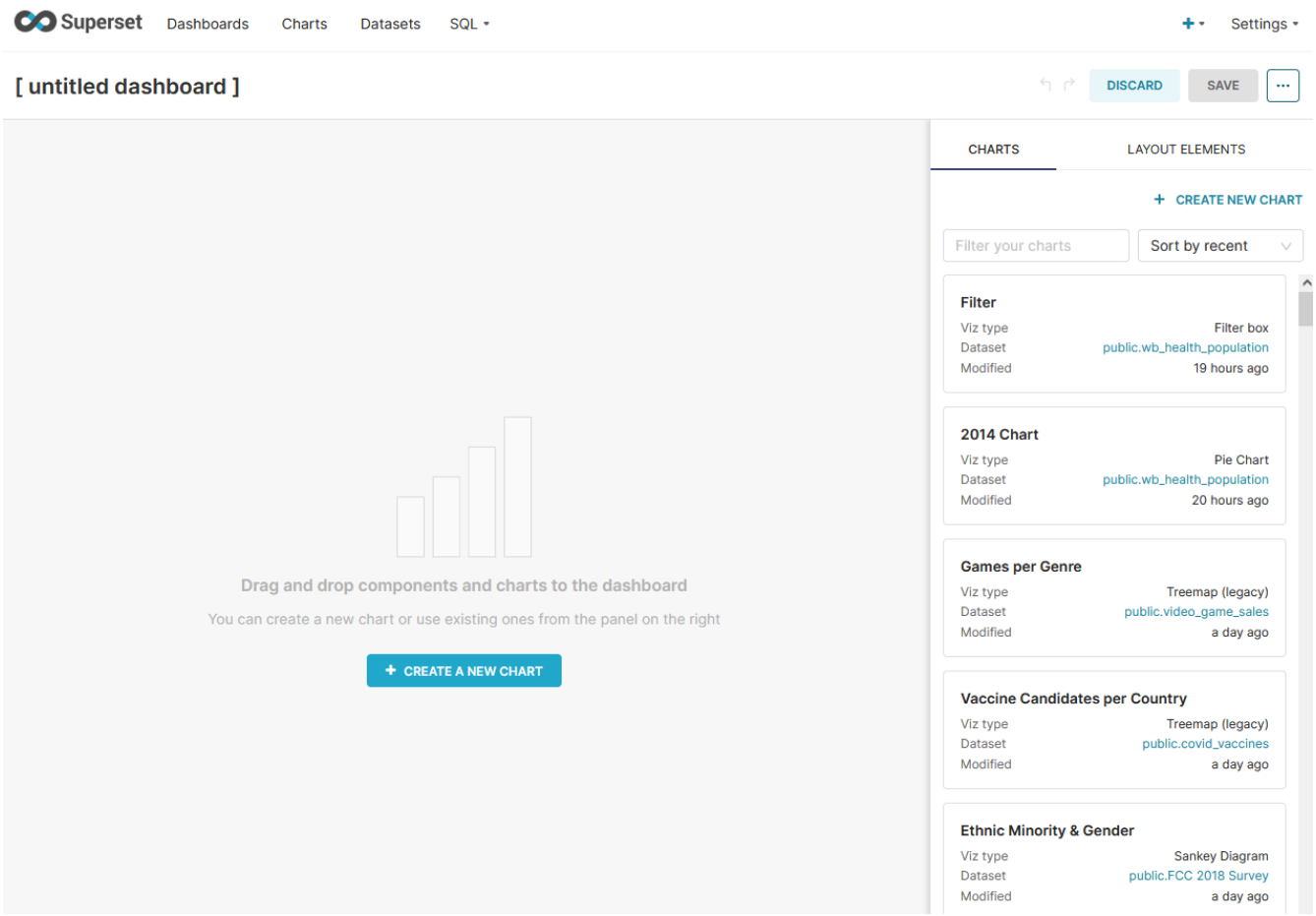

Abbildung 6. Ansicht für ein neues Dashboard

Im Bereich **Untitled dashboard** kann dann dem Dashboard ein Name gegeben werden.

Bei einem neu erstellten Dashboard kann man direkt zu editieren anfangen. In jedem anderen Fall musst du oben rechts auf *Edit dashboard* klicken. Momentan ist dein Dashboard noch leer, jedoch kannst du dieses einfach per Drag & Drop füllen. Die erste Komponente musst du nach oben zum Rand ziehen. Wenn eine Komponente platziert werden kann, wird dies durch eine blaue Linie signalisiert, die zugleich anzeigt, wie/wo die Komponente platziert wird. Zuerst klickst du dafür auf die *Edit dashboard*-Schaltfläche wodurch alle Komponenten angezeigt werden, die du hinzufügen kannst.

An Elementen unterscheidet man im Wesentlichen zwischen Layout-Elementen und Charts. Layout-

Elemente sind dabei vorgefertigte Elemente, welche rein der Gestaltung des Dashboards dienen, also nicht mit den Daten interagieren. Die Charts hingegen sind entweder Diagramme oder Filter, also Elemente, welche die Daten entweder filtern oder visuell Darstellen. In diesem Bereich findet man auch alle User-Erstellten Elemente, z.B. die in der vorherigen Aufgabe erstellten Elemente.

Die Elemente können alle interaktiv umhergeschoben sowie vergössert und verkleinert werden.

Aufgabe 2: Ein Filter für das Dashboard

Als nächstes möchten wir ein Filter für den Datensatz erstellen.

Dafür müssen in einem erstellten Dashboard den Button **[→ |]** gedrückt werden und unter **dem Zahnrad › Add or edit filters** die Rahmenbedingungen des Filters definiert werden.

Im Fall der Aufgabe wird unter **Filter name** ein treffender Name z.B. "Country Name" ausgewählt, und unter **Column** die Spalte “country_name”

Sobald der Filter gespeichert wurde, können unter **Filters › `Country Name`** einzelne Länder gefiltert werden.

Wenn als Column die Spalte “year” gewählt wird, können die zehn bevölkerungsreichsten Länder zu einem bestimmten Zeitpunkt angezeigt werden. Dabei ist es Sinnvoll, bei der Erstellung der Filter die Box unter **Filter Settings › “Can select multiple values”** abzuwählen.

Aufgabe 3: Ein Dashboard erstellen

Erstelle nun ein Dashboard, das der Abbildung 7 entspricht.

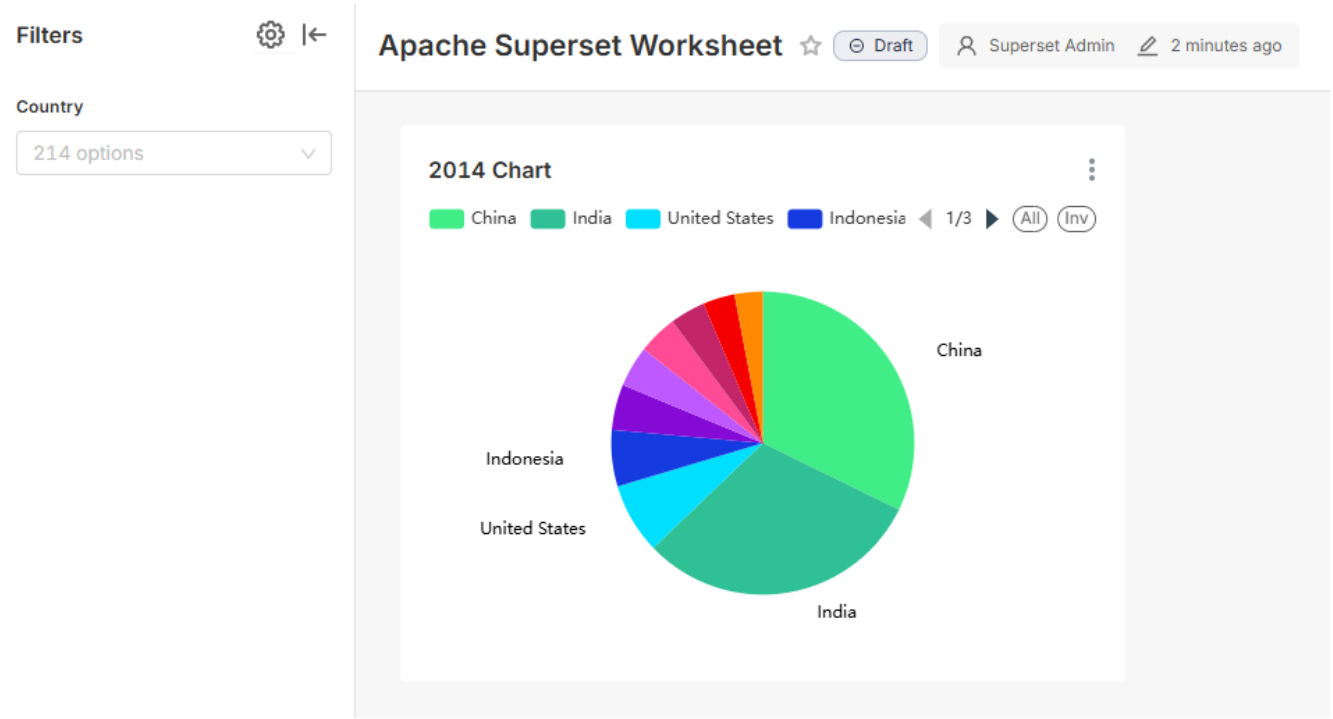

Abbildung 7. Hier nochmals das Dashboard als Ergebnis dieses Arbeitsblatts mit dem Pie-Chart und dem Filter

Dashboards teilen

Sobald du mit deinem Dashboard zufrieden bist, kannst du es publizieren und/oder teilen.

Um ein Dashboard zu teilen, musst du hinter der **EDIT DASHBOARD**-Schaltfläche auf das Dropdown-Menü und dort auf **… › Share › Copy Permalink to Clipboard**. Die URL kannst du nun einer anderen Person schicken, jedoch muss der Account dieser Person der entsprechenden Rolle zugewiesen sein.

Abschluss

Geschafft! Du solltest nun ein Dashboard zu den Weltbank-Daten haben, das du anderen zeigen kannst.

Tipp zum Filter: In einem Zeit-Filter unter *Custom* ist es möglich, direkt Jahreszahlen zu schreiben. Das Datum ist dann automatisch der erste Januar.

Wer mehr über Apache Superset erfahren will, dem seien die ausführlicheren Informationsblätter "Einführung in Apache Superset (7 Charts)" und "Apache Superset für Fortgeschrittene" auf OpenSchoolMaps empfohlen.

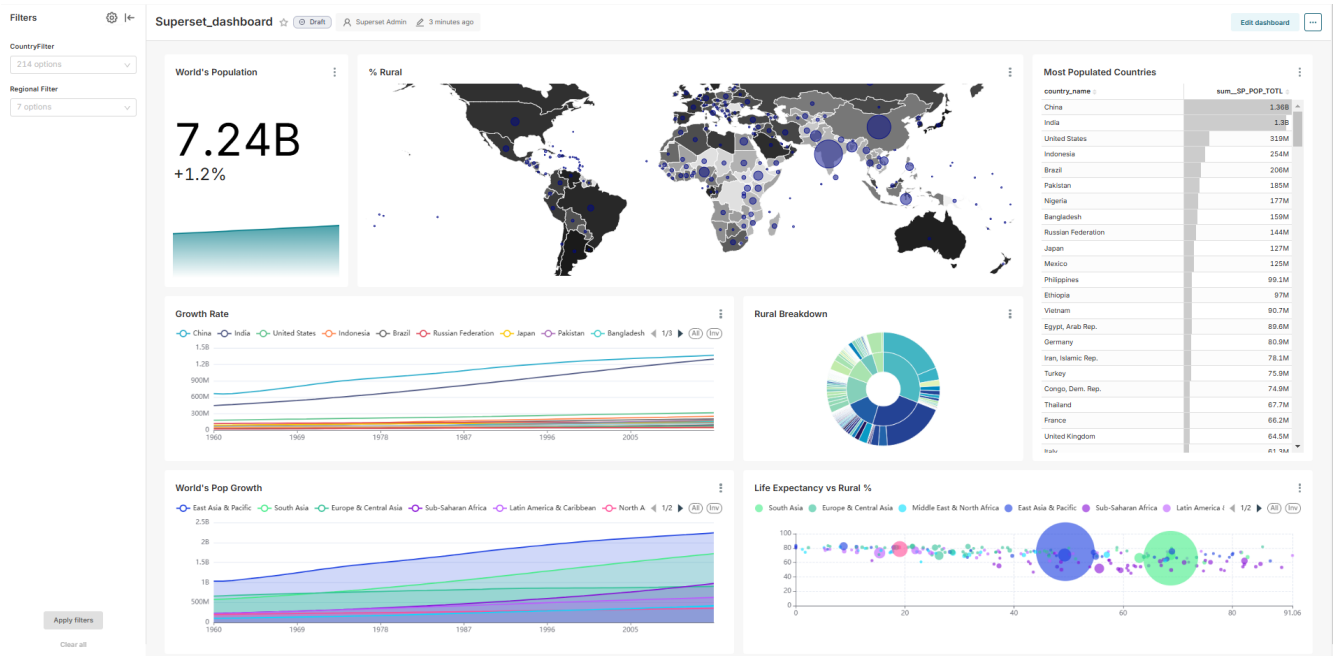

Abbildung 8. Diagramme, die in der Einführung in Apache Superset (7 Charts) vorgestellt werden.

Noch Fragen? Siehe "Kontakt" auf OpenSchoolMaps!