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 0.34 von Apache Superset, der im September 2019 freigegeben wurde.

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. Diese scheint leider sehr dünn und nicht nachgeführt zu sein: Sie verwendet veraltete Begriffe ("Slices" statt "Charts") und das darin enthaltene Tutorial setzt Wetterdaten voraus, die nicht vorinstalliert sind. Der nützlichste Teil der Dokumentation ist das kleine FAQ Frequently Asked Questions (Häufig gestellte Fragen): https://superset.incubator.apache.org/faq.html .

Benutzerkonto erstellen/anmelden

Apache Superset ist eine Webapplikation und für den Zugang wird je nach Server ein Konto benötigt. Wenn du bereits Zugang zu einer Apache-Superset-Instanz hast, dann melde dich dort an.

Apache Superset kennt verschiedene Benutzer-Rollen, die bestimmte Rechte haben, um Daten zu verändern oder Funktionen aufzurufen. In diesem Arbeitsblatt wird angenommen, dass du Benutzerrechte erhalten hast, die der "Gamma"-Rolle entsprechen, d.h. du kannst Charts und Dashboards erstellen.

Konzepte und Begriffe

Nach dem Anmelden bei einer Apache Superset-Applikation (siehe vorhergehendes Kapitel) siehst du das Menü oben und unter anderem die Menüpunkte *Sources*, *Charts* und *Dashboards*.

Hier einige Erläuterungen zu den Konzepten hinter Apache Superset:

- Datasource (Datenquelle): Eine "Source" kann zum Beispiel durch das Hochladen einer CSV-Datei oder durch die Erstellung von Views mittels SQL-Abfragen erzeugt werden.
- Datenbank: Eine Verbindung (Connection) zu einem Datenbanksystem, das Tabellen und Views enthält.
- Chart (Diagramm), früher "Slice" genannt: Ein "Chart" ist eine Liste, ein Diagramm oder eine Webkarte. Superset kennt zurzeit 48 verschiedene Charts, wobei sieben davon interaktive Webkarten ("Map Charts") sind.
- Dashboards: Eine interaktive Webseite, auf der interaktive Charts präsentiert werden.
- Metrics: "Metrics", manchmal auch "Measurements" genannt, sind numerische Kennzahlen. Sie werden v.a. in den Charts erwähnt und verlangt.
- Records: 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.

Im Menü "Sources" von Apache Superset und in der Orginal-Dokumentation von Apache Superset wird Druid als Datenbanksystem erwähnt. Alle auf Druid bezogenen Informationen im Menü und in der Dokumentation kannst du ignorieren.

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)"). Hilfreich, um Daten zu bereinigen ("Data Wrangling"), kann hierbei auch z.B. OpenRefine sein.

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:

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

Abbildung 1 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	population_total
Europe & Central Asia	Switzerland	1960	5327827
Europe & Central Asia	Switzerland	1961	5434294
Europe & Central Asia	Switzerland	1962	5573815
Europe & Central Asia	Switzerland	1963	5694247
Europe & Central Asia	Switzerland	1964	5789228
Europe & Central Asia	Switzerland	1965	5856472
Europe & Central Asia	Switzerland	1966	5918002
Europe & Central Asia	Switzerland	1967	5991785
Europe & Central Asia	Switzerland	1968	6067714
Europe & Central Asia	Switzerland	1969	6136387
Europe & Central Asia	Switzerland	1970	6180877
Europe & Central Asia	Switzerland	1971	6213399
Europe & Central Asia	Switzerland	1972	6260956
Europe & Central Asia	Switzerland	1973	6307347
Europe & Central Asia	Switzerland	1974	6341405
Europe & Central Asia	Switzerland	1975	6338632
Europe & Central Asia	Switzerland	1976	6302504

Abbildung 1. Daten der Schweiz von 1960 - 1976 aus der Tabelle wb_health_population.

Charts (Diagramme)

Apache Superset bietet derzeit 40 Charts an. Im Anhang findest du weitere Beispieldiagramme zur Tabelle 'wb_health_population', die in weiterführenden Arbeitsblättern vorgestellt werden (Siehe Kapitel "Abschluss").

Am Ende dieser Anleitung wirst du einen Chart und einen Filter erstellt haben. (Abbildung 2)

- Links befindet sich ein Filter. Damit kann man die Datensätze auf bestimmte Regionen und Länder beschränken.
- Rechts befindet sich ein Kreis-Diagramm (Pie Chart), welches die zehn bevölkerungsreichsten Länder und deren Einwohnerzahlen (Metric) anzeigt.

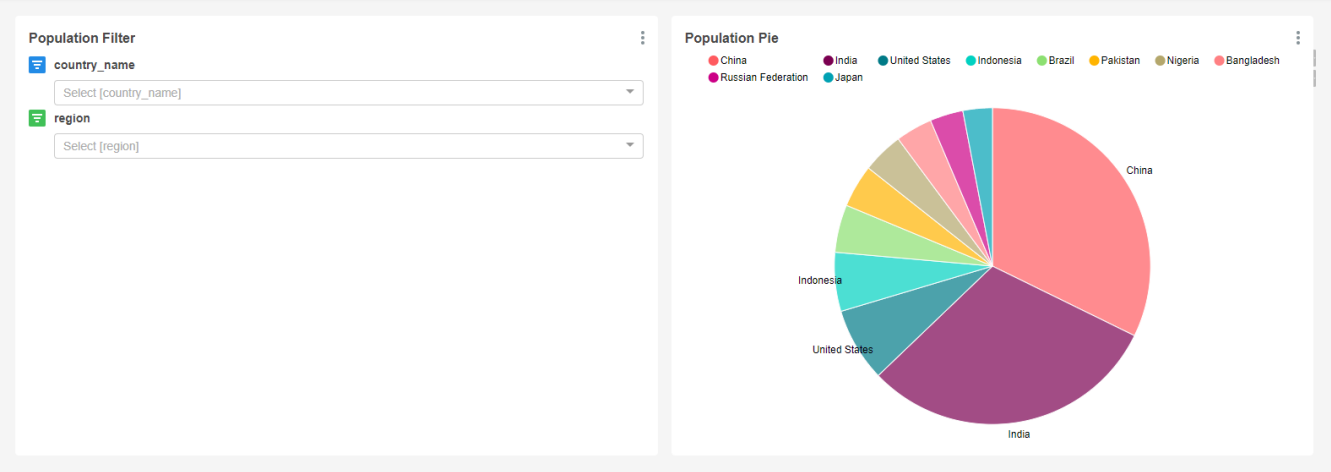


Abbildung 2. Das Dashboard mit dem Kreis-Diagramm und dem Filter.

Wenn die Charts während der Erstellung nicht richtig angezeigt werden, kann das zum Teil behoben werden, indem man die Website neu lädt (F5).

Aufgabe 1: Mein erster Chart

Als erstes musst du eine Datenquelle (Source) wählen. Unter Sources können die verschiedenen Tabellen eingesehen werden. Hierbei kann man auf die Lupe klicken, um genauere Informationen über die Tabelle zu erhalten, wie z.B. welche Spalten die Tabelle besitzt und welche Datentypen dort gespeichert werden können. Zudem kann man unter *Columns* einstellen, ob die Spalte zeitlich oder filterbar ist und ob nach ihr gruppiert werden darf.

Um eine Tabelle abzufragen, muss man unter "Sources" den Menüpunkt "Tables" wählen und dann auf den Namen der Tabelle klicken. In diesem Beispiel ist das `wb_health_population`. Man kann 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.

Durch das Auswählen einer Tabelle öffnet sich direkt ein neues Fenster bzw. ein neuer Browser Tab, worin man nun einen Chart erstellen kann. Dort steht nun bei Datasource die gewählte Tabelle `wb_health_population`.

Direkt darunter, bei *Visualization Type* (Chart-Typ) kann anstatt *Table* "Pie Chart" (Kreis-Diagramm) ausgewählt werden.

Als erstes musst du unter *Time* die Zeitspanne anpassen, da die Daten der Tabelle `wb_health_population` im Jahre 1960 beginnen und im Jahr 2014 aufhören. Da wir in unserem Beispiel nur das Jahr 2014 betrachten wollen, verwenden wir einen *Custom*-Filter von `2014-01-01` – `2014-12-31`.

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.

Um z.B. das Jahr 2014 auszuwählen, kann man auch `2013` bis `2014` eintippen.

Unter *GROUP BY* kannst du jetzt deine erste Abfrage definieren, indem du bei *Metric* eine Spalte wie `SP_POP_TOTL` (Bevölkerung Total) auswählst und davon die Summe nimmst. Dies kannst du tun, indem du auf das Textfeld unter *Metric* klickst und ein *aggregate* (das sind die Einträge, vor welchen ***AGG*** steht) gefolgt von der entsprechenden Spalte wählst. Die Spalte *Metrics* muss immer ausgefüllt sein. Sie wird auch bei allen Darstellungen verwendet, um zu bestimmen, wie viel Gewicht diese Zeile hat. Bei unserem Pie-Chart ist es die Grösse des Abschnittes.

Unter *GROUP BY → Group by* kannst du die Abfrage noch weiter unterteilen. *Group by* wird genutzt, um ein Total nach einem bestimmten Attribut aufzuteilen. Wenn z.B. die Summe der Weltbevölkerung nach Ländernamen gruppiert wird, erhält man die Bevölkerung pro Land. Wähle darum hier das Attribut `country_name` aus. Das davorgestellte "ABC" zeigt den Datentyp dieses Attributes an, was an dieser Stelle noch nicht wichtig ist.

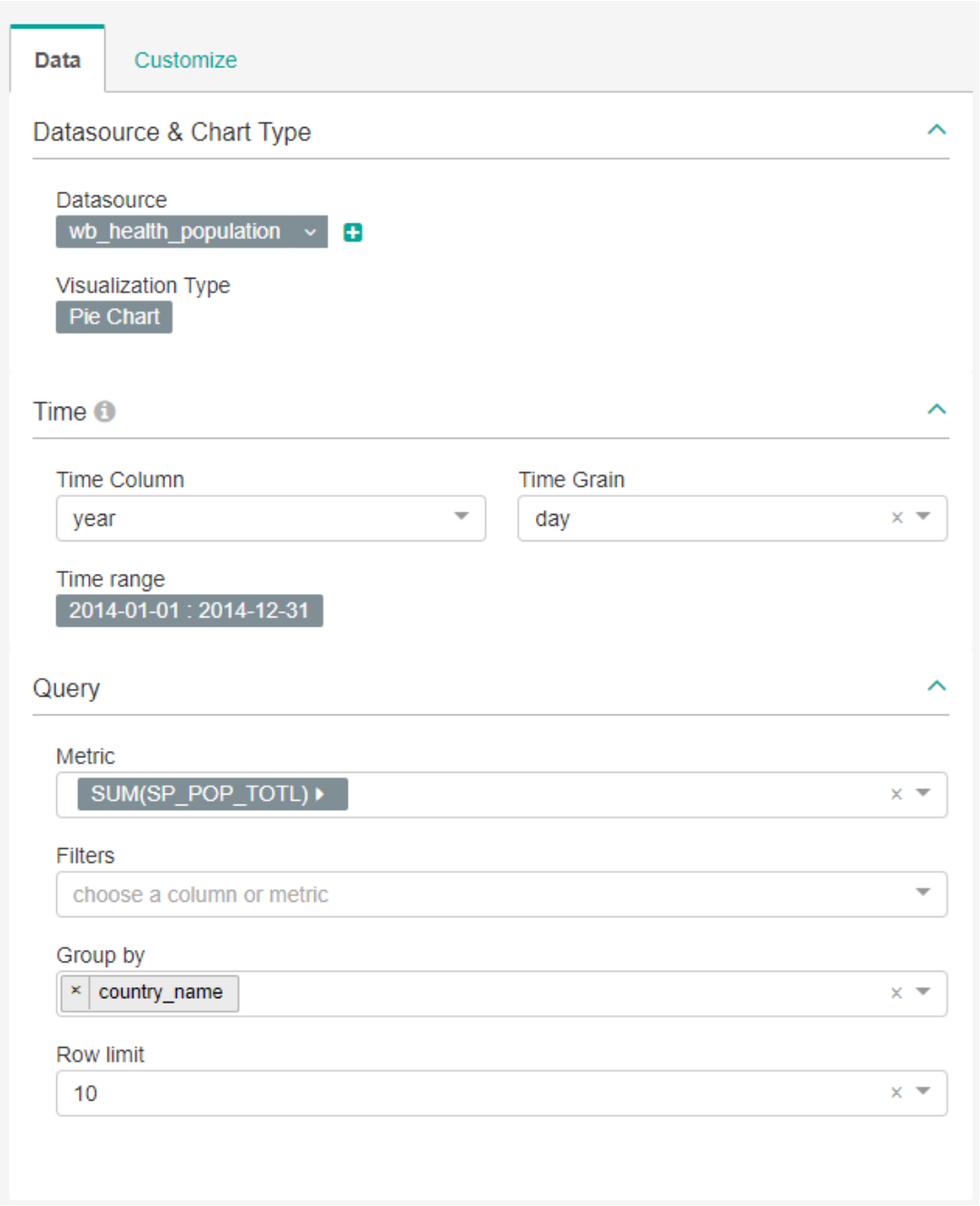
Unter *Row Limit* kannst du das Ergebnis auf eine bestimmte Anzahl einträge beschränken. Wenn du z.B. ein *Row Limit* von 10 setzt, werden nur die 10 bevölkerungsreichsten Länder angezeigt.

Wenn du jetzt *Run Query* drückst, wird die Abfrage ausgeführt. Sie zeigt in einem Kreis-Diagramm die zehn bevölkerungsreichsten Länder im Jahre 2014.

Speichere nun diesen Chart als dein erstes Ergebnis in Apache Superset, z.B. mit dem Namen "Population Pie Chart". Du wirst diesen Chart später wieder brauchen.

Lösung

Die Abfrage müsste wie folgt aussehen:



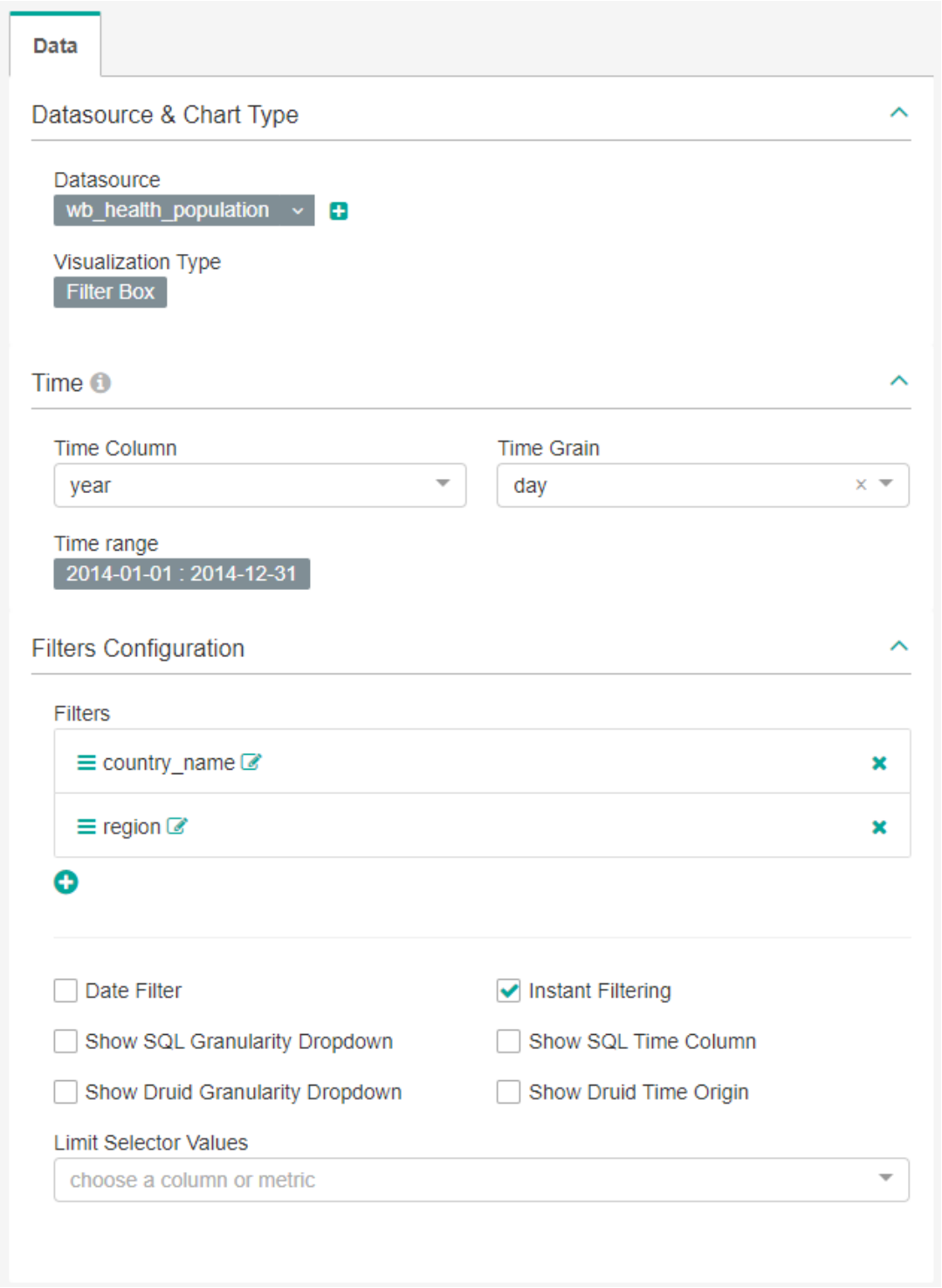
Aufgabe 2: Filter Box für das Dashboard

Jetzt, da wir einen Chart haben, wäre ein Filter praktisch, den man nachher im Dashboard anwenden kann. Dafür muss man nur die *Filter Box* als *Visualization Type* auswählen und unter *Filters* passende Spalten auswählen z.B. `region` und `country_name`. Zudem kann beim Filter die Option *Date Filter* abgewählt werden.

Diesen Filter muss man natürlich auch abspeichern, um ihn später benutzen zu können.

Lösung

Die Abfrage müsste wie folgt aussehen:



Charts zu einem Dashboard anordnen

Unter Dashboards kannst du nun dein eigenes Dashboard erstellen, indem du oben rechts auf das + drückst. In diesem Dashboard können nun dein Chart und dein Filter präsentiert werden.

Beim Erstellen eines Dashboards kann man einiges einstellen, jedoch reicht es oft, nur das Feld *Title* auszufüllen.

Um mit dem Editieren zu beginnen, 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 den *Edit dashboard*-Knopf wodurch alle Komponenten angezeigt werden, die du hinzufügen kannst.

- *Tabs* sind wie beim Browser selber zu verstehen und können alles Mögliche beinhalten.
- *Rows* sowie *Columns* können gebraucht werden, um einzelne andere Komponenten zu verbinden. Wenn Komponenten verbunden sind, kann man den Platz zwischen ihnen auch weiss färben. Durch das Hovern über den verbundenen Komponenten erscheint ein Knopf, der diese Option anzeigt.
- Um die verbundenen Komponenten wieder in Gruppen zu unterteilen kann man *Divider* verwenden. Diese stellen einen Strich dar.
- *Header* können verwendet werden, um den einzelnen Komponenten eine Überschrift zu geben. *Markdown* kann zum Verfassen eines beschreibenden Textes zu einem Chart benutzt werden.
- Unter *Your charts & filters* sind alle Charts zu finden, die man bisher erstellt hat.

Nach dem Hinzufügen einer Komponente kann deren Grösse durch Anklicken der unteren rechten Ecke angepasst werden.

Aufgabe 3: Ein Dashboard erstellen

Erstelle nun ein Dashboard, das der Abbildung 3 entspricht.

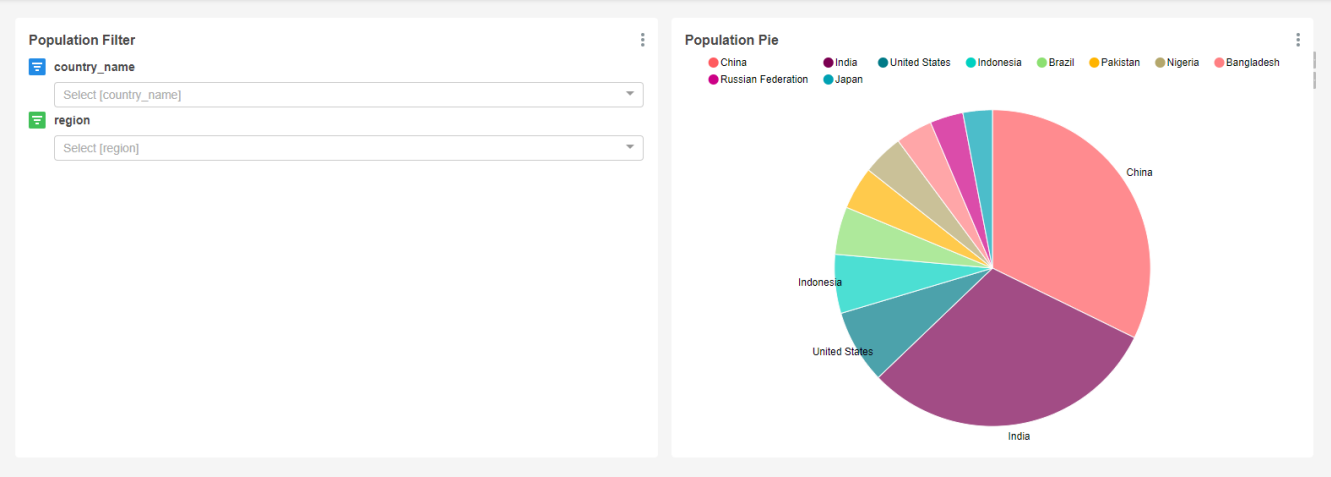


Abbildung 3. Hier nochmals das Dashboard als Ergebnis dieses Arbeitsblatts mit dem Pie-Chart und dem Filter

Dashboards mit anderen (übers Web) teilen

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

Um ein Dashboard zu teilen, musst du hinter dem *Edit dashboard*-Knopf auf das Dropdown-Menü und dort auf *Share dashboard* klicken. Die URL kannst du nun einer anderen Person schicken, jedoch muss diese Person (d.h. deren Rolle) auch Zugriff auf die Daten haben, andernfalls wird nur

ein Fehler (Error…) angezeigt.

Um ein Dashboard zu publizieren, musst du lediglich rechts neben dem Namen des Dashboards auf "Draft" drücken, damit "Published" angezeigt wird. Ein publiziertes Dashboard ist für andere User unter Dashboards sichtbar (sofern diese User Zugriff auf die entsprechenden Datasources haben). Ein Dashboard, welches als "Draft" markiert ist, ist zwar nicht für andere User sichtbar (Ausnahme: hochrangige User wie z.B. Admins), dennoch kann man mittels URL direkt drauf zugreifen.

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.

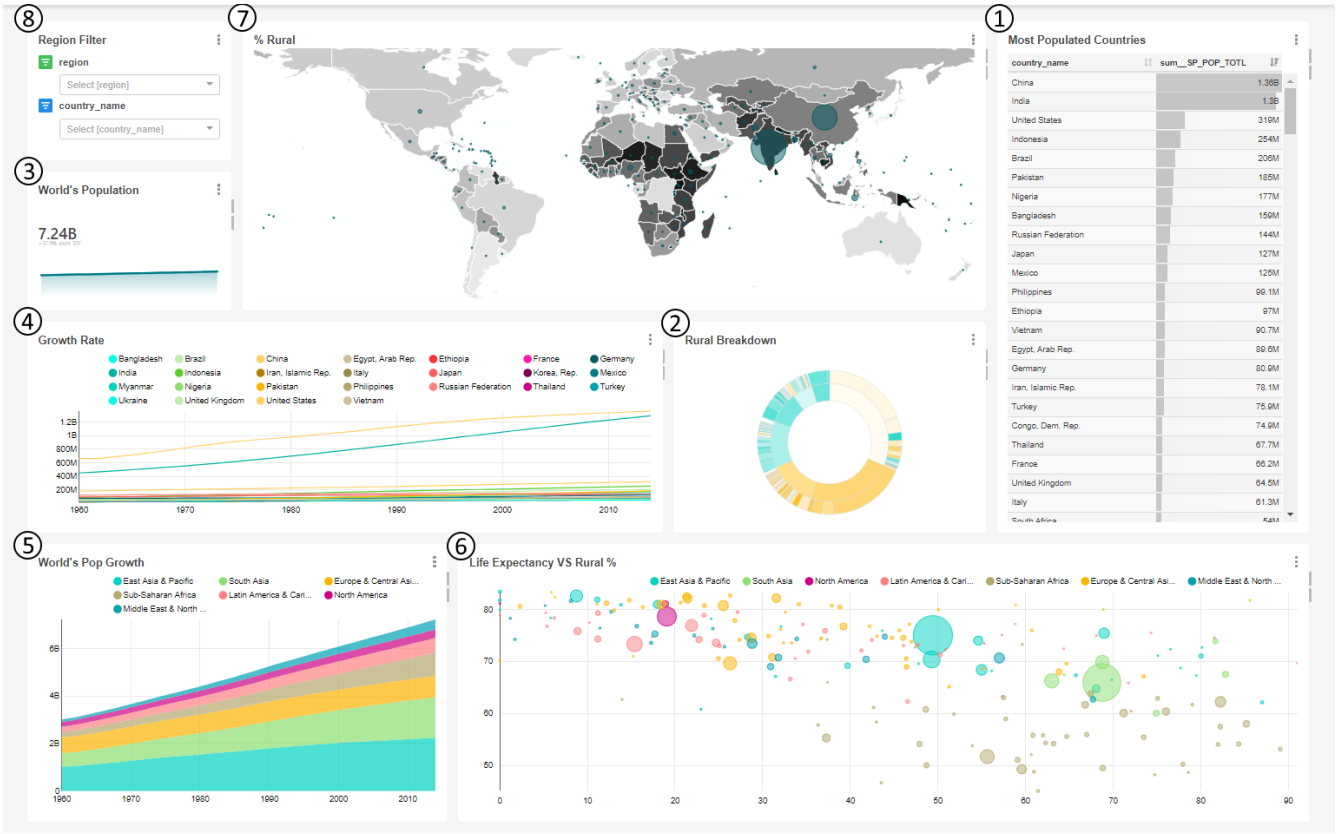


Abbildung 4. Diagramme, die in der Einführung in Apache Superset (7 Charts) vorgestellt werden.

Zu Apache Superset gibt es auch Kurse, u.a. am Geometa Lab HSR.

Gerne nehmen wir Rückmeldungen entgegen, siehe OpenSchoolMaps > *Weitere Unterrichtsideen*.