Vergleich der Visualisierungsmöglichkeiten in Qlik und Python am Beispiel Human Development

Philippe Denig, Jakob Dietrich, Luis Rastetter

9. Mai 2024

Inhaltsverzeichnis

1 Einleitung

Datenvisualisierung ist ein wichtiger Bestandteil der modernen Datenanalyse. Visualisierungen ermöglichen es, komplexe Informationen effektiv zu kommunizieren, sodass der Empfänger diese schnell erfassen kann. Dadurch spielt Datenvisualisierung in der Wirtschaft und Wissenschaft, aber auch im Privatleben eine entscheidende Rolle, um Erkenntnisse zu gewinnen und Zusammenhänge zu verstehen. Insbesondere im Geschäftsumfeld können gute Visualisierungen dazu beitragen, wichtige Entscheidungen schnell und gut treffen zu können.

Angesichts der Bedeutung der Datenvisualisierung gibt es eine Vielzahl von Tools und Technologien, die mit unterschiedlichen Ansätzen und Funktionen versuchen bestmögliche Ergebnisse zu liefern. In dieser Arbeit wird sich mit den der Programmiersprache Python im Datenvisualisierungskontext und der Business Intelligence Plattform Qlik beschäftigt und es soll herausgestellt werden, welches Tool sich besser für welche Anwendungszwecke eignet.

Um einen fundierten Vergleich zu ermöglichen, werden erst beide Tools kurz vorgestellt. Anschließend erfolgt eine systematische Evaluierung anhand verschiedener Aspekte, die sich an den einzelnen Schritten des CRISP-DM Verfahrens orientieren. Das CRISP-DM Verfahren besteht aus den Phasen Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation und Deployment. Es wird sich jedoch überwiegend auf die Phasen konzentriert, die durch die Tools beeinflusst werden: Data Understanding, Data Preparation und Modelling. Business Understanding, Evaluation und Deployment betreffen primär strategische Aspekte und werden somit nicht direkt von den Funktionen der Tools beeinflusst. Um einen konkreten Anwendungsfall zu illustrieren, wird das Thema Human Development als Beispiel herangezogen. Dadurch wird ein realer Kontext geboten, in dem die Wirksamkeit der Tools bei der Analyse komplexer sozialer und wirtschaftlicher Daten bewertet werden kann. Darüber hinaus wird eine zusätzliche Kategorie SSonstigeseëingeführt, die sich mit anderen, nicht in den CRISP-DM Prozess einordbaren Aspekten wie Kostenstruktur oder Dokumentation und Support beschäftigt. Dadurch wird eine umfassendere Einschätzung in dem Gesamtkontext der Data Science und Business Intelligence Aktivitäten eines Unternehmens ermöglicht.

Das Ziel ist es, ein tiefgreifendes Verständnis für die Stärken und Schwächen

von Python und Qlik im Rahmen der praxisrelevanten Abschnitte des CRISP-DM Verfahrens zu erlangen. Auf Basis dieser methodischen Herangehensweise kann in Zukunft die Entscheidung des am besten für einen spezifischen Anwendungsfall geeigneten Visualisierungstool einfacher getroffen werden.

1.1 Was ist Qlik

Qlik ist eine der führenden Plattformen im Bereich der Business Intelligence und wurde speziell für interaktive Datenanalyse und Datenvisualisierung konzipiert. Das Qlik Universum umfasst eine breite Produktpalette, durch die eine umfassende BI-Lösung zur datengetriebenen Entscheidungsunterstützung für Unternehmen verschiedener Größen und Branchen geboten wird. Im Zentrum stehen die beiden Datenvisualisierung und Datenalyse Tools Qlik View und Qlik Sense, diese werden dann durch weitere Anwendungen wie z.B. Qlik Data für Datenmanagement unterstützt werden. Qlik View dient primär der Analytik und ist auf die Entwicklung von individuellen Anwendungen für die Datenanalyse ausgerichtet. Es ermöglicht Benutzern, komplexe Daten aus verschiedenen Quellen zu kombinieren, zu transformieren und zu visualisieren, um Einblicke zu gewinnen und fundierte Entscheidungen zu treffen. Qlik Sense hingegen wurde entwickelt, um eine breitere Benutzerbasis anzusprechen und bietet eine benutzerfreundliche, self-service-orientierte Plattform für die Datenvisualisierung und -analyse. Mit Qlik Sense können Benutzer interaktive Dashboards und Berichte erstellen, um Daten auf intuitive Weise zu erkunden und zu verstehen.

In dieser Arbeit werden ausschließlich die Qlik Coud Services in der "Qlik Sense Business - International Version"betrachtet, da diese Version einen 30-tägigen Testzeitraum anbietet, mit dem unverbindlich und kostenlos ein umfassender Einblick in die Plattform gegeben wird. Außerdem bietet Qlik Sense im Vergleich zu Qlik View ein flexibleres Arbeitsumfeld und eine simplere Bedienung und steht somit mehr für moderne BI-Tools und grenzt sich weiter von Python ab.

1.2 Was ist Python

Python ist eine der am weitesten verbreiteten Programmiersprachne und spielt eine zentrale Rolle im Feld der Datenvisualisierung. Dank vieler starter Bibliotheken in

diesem Bereich wie z.B. Matplotlib, Seaborn und Plotly bietet Python eine schier endlose Bandbreite and Möglichkeiten. Matplotlib bietet eine solide Grundlage für Diagramme und Graphiken, Seaborn bietet einen mehr auf Statistik basierneden Ansatz und mit Plotly können die Darstellungen weiter verfeinert werden, indem z.B. interaktive Elemente hinzugefügt werden. Außerhalb des tatsächlichen Visualisierungsprozesses kommen noch weitere Datenverarbeitungswerkzeuge wie Numpy und Pandas. Außerdem hat Python eine sehr große Community und es existieren zahlreiche kostenlose Lernuntertützungen. Insgesamt bietet Python eine flexible Basis und in Kombination mit den entsprechenden Bibliotheken ist fast alles möglich.

1.3 Datensätze

Da der Toolvergleich am Beispiel Human Development durchgeführt wird, werden hier kurz die verwendeten Datensätze und deren Quellen vorgestellt. Human Development ist ein sehr breites Themengebiet, daher werden verschiedene Aspekte wie z.B. Lebenserwartung, Geburtenrate und Todesfälle durch Naturkatastrophen betrachtet.

Dazu werden insgesamt 10 Datensätze genutzt, 7 davon stammen von UNData. UNData ist eine Datensammlung der Vereinten Nationen über verschiedenste Themen, unter anderem Humand Development. Daher sind die in ihrem Format auch sehr ähnlich und bestehen immer aus einer Spalte Land, einer Spalte Jahr und dann dem Wert des jeweiligen Datensatzes. Da somit immer nur eine Datenspalte pro Datensatz vorliegt, werden 7 Datensätze benötigt. Die Daten stammen primär von den einzelnen Unterorganisationen der UN wie beispielsweise der WHO und sind somit sehr vertrauenswürdig. Einer der übrigen drei Datensätze über Selbstmordraten stammt von der WHO direkt, da er nicht über UNData veröffentlicht wurde und ist somit auch als sehr vertrauenswürdig einzustufen. Die restlichen beiden Datensätze stammen von Kaggle, einem Webportal auf dem Datensätze veröffentlicht werden können. Zum einen stammt ein Datensatz über Naturkatastrophen von Kaggle. Die ursprüngliche Quelle des Datensatzes liegt bei der NASA und wurde größtenteils mit dem Satellitenprogramm EOSDIS und dessen Vorgängern gesammelt. Grundsätzlich ist dieser Datensatz also ebenfalls sehr vertrauenswürdig, es gilt aber zu beachten, dass die Daten zwar bis 1900 zurückreichen, aber bis 1970 insbesondere kleine Naturkatastrophen fehlen. Der letzte Datensatz ist der World Happyness

Report (WHR) aus dem Jahr 2021. Der WHR bewertet das Glücksempfinden der Menschen verschiedener Länder und wird von einem internationalen Team aus Wissenschaftlern zusammengestellt und somit ist auch er eine seriöse Quelle darstellt. Es gilt aber zu beachten, dass der hier verwendete finale Wert des Glücksempfindens auf Basis vieler Faktoren berechnet wird und nur ein Teil der Bewertung auf Umfragen an den Bürgern der Länder basiert.

Durch die Verwendung solch qualitativ hochwertiger und vertrauenswürdiger Datensätze wird sichergestellt, dass die zum Thema Human Development gewonnenen Erkenntnisse auch aussagekräftig sind und dass die Visualisierungen mit den beiden Tools in einer realistischen Situation getestet und bewertet werden kann.

2 Vergleich zwischen Qlik und Python

2.1 Einleitung Vergleich

Wie schon in der Einleitung erwähnt, erfolgt der Vergleich anhand der relevanten Phasen des CRISP-DM Prozesses mit der zusätzlichen Kategorie Sonstiges. In den einzelnen Phasen werden Python und Qlik anhand verschiedener für diese Phase relevanten Kriterien evaluiert und es wird ein Bezug zu dem Visualisierungsprozess an dem Thema Human Development hergestellt, der insbesondere durch die große Anzahl an verschiedenen Datensätzen eine besondere Herausforderung darstellt. Im folgenden Vergleich wird davon ausgegangen, dass er Anwender Grundkenntnisse in der Programmierung hat, da ansonsten eine Verwendung von Python ohne vorherigen Lernprozess nicht in Frage kommt, da es zwar im Vergleich zu anderen Programmiersprachen intuitiv ist, aber dennoch nicht ohne Programmierkenntnisse bedienbar ist. Die erforderliche Programmierkenntnisse stellen natürlichen einen großen Minuspunkt dar, auf diesen wird dann wieder im Fazit eingegangen.

2.2 Data Understanding

Im ersten Schritt des Toolvergleichs wird sich auf das Data Understanding fokussiert, bei dem es darum geht, die Daten im entsprechenden Tool hinzuzufügen und eine umfassende Übersicht über die verfügbaren Datenquellen, deren Qualität, Struktur und potenzielle Einschränkungen zu gewinnen.

In Qlik lassen sich neue Datensätze in Form von einzelnen Dateien sehr schnell per Drag and Drop hinzufügen und es werden alle gängigen Dateiformate bis auf JSON unterstützt. Bei JSON Datein ist eine Vorverarbeitung notwendig. Das Hinzufügen der Datein in Python dauert bedeutend länger, da sie erst per Codezeile eingelesen werden müssen und das Schreiben dieser einige Zeit braucht, dafür existieren keinerlei Einschränkungen bezüglich des Dateiformats. Der Pozess des Daten hinzufügens in Qlik ist sehr intuitiv und gut beschrieben, sodass schon bei der ersten Nutzung alles schnell ohne Probleme funktioniert hat. Auch große Datensätze stellen kein Problem da. Laut Qlik liegt die maximale Dateigröße bei 100 Gigabyte. Die Geschwindigkeit des Uploads hängt von der Uploadgeschwindigkeit des Users ab, da die hier genutze Qlik Version mit einer Cloud arbeitet. Python kann sowohl lokal, als auch in einer Cloud verwendet werden. Bei lokaler Nutzung gibt es keine Verzögerungen durch den Upload, bei Nutzung in einer Cloud hängt es von der Uploadgeschwindigkeit des Nutzers ab. In Python ist eine Nutzung von APIs mit entsprechenden Bibliotheken wie z.B. Request relativ einfach möglich. Auch in Qulik können APIs verwendet werden, aber die Anzahl der APIs ist insbesondere in der Standart Version etwas begrenzt.

In dem hier verwendeten Beispiel Human Development wird eine größere Anzahl an Datensätzen benötigt. Qlik bietet eine sehr gute Übersicht darüber, welche Datensätze hochgeladen sind. Die Datensätze lassen sich in Raster- und Listenansicht anzeigen und nach Kriterien wie ”recently usedßorteiren, außerdem lassen sich Favoriten festlegen. In Python kann man durch eine übersichtliche Ordnerstruktur im Kombination mit einer guten IDE den Überblick über die verschidenen Dateien bewahren.

Der nächste wichtige Schritt des Data Understanding ist die Datenexploration. Das Tool sollte dem Nutzer Möglichkeiten bieten, einen groben Überblick über den Datensatz zu bekommen, insbesondere NaN-Werte erkennen spielt hier eine sehr große Rolle.

In Python lässt sich Datenexploration insbesondere mit Pandas sehr gut durchführen. Durch die Nutzung eines Panda Dataframes werden dem Nutzer sehr viele Befehle wie head(), info(), describe() zur Verfügung gestellt. Mit diesen lässt sich ein sehr guter Überblick über die Struktur des Datensatzes erhalten. Außerdem gibt es die Funktion isna(), mit der NaN-Werte identifiziert werden können. In Qlik kann

sich über die allgemeine Struktur des Datensatzes in dem Overview Reiter eine gute Übersicht gebildet werden. Für weitere Details kann in den Fields Reiter gewechselt werden. In diesem sind Informationen über die Einzelnen Spalten eines Datensatzes zu finden und es werden Visualisierungen zur Verteilung der Werte angezeigt. Zudem ist hier in Einblick der ersten 10 Zeilen in Tabellenform zu finden und eine Auflistung der einzelnen Spalten inklusive Datentyp.

Bisher liegen in Qlik die Daten nur als einzelne in die Cloud hochgeladene Dateien vor, mit ihnen kann aber noch nichts bis auf die Datenexploration der einzelnen Datensätze gemacht werden. Um weiter mit den Daten zu arbeiten muss eine App, in Qlik ist eine App ein Arbeitsdokument, das eine analytische Anwendung darstellt, erstellt werden und die Daten müssen hinzugefügt werden. Das Hinzufügen der Daten geht sehr einfach, der Benutzer kann aus den Daten die er in Qlik hochgeladen hat auswählen, welche er der App hinzufügen möchte. Es ist auch möglich nur einzelne Spalten aus Datensätzen auszuwählen. Dieser Prozess hat in der Basisversion ein Limit von 1260 Megabyte und dauert bei größeren Datensätzen bishin zu wenigen Minuten. Nach dem Datenupload werden sofort, ohne das weitere Aktionen des Benutzers erfolgen müssen Vorschläge unterbreitet, wo sich die einzelnen Datensätze miteinander verknüpfen lassen. Danach bedindet man sich direkt im "Data Manager", in dem man per Drack and Drop die einzelnen Datensätze entknüpfen und verknüpfen kann. Die zusammengehörenden Spalten werden automatisch festgestellt. Falls dies nicht möglich ist, kann manuell gewählt werden, welche Spalten verknüpft werden sollen. Außerdem können hier Datensätze in zwei Teile aufgetrennt werden. Im Datenmanager existiert neben der Verknüfungsperspektive die tabellenperspektive, hier lassen sich ein einzelne Werte wie z.B. NaN-Werte suchen. Zur besseren Visualisierung der Verknüpfungen kann auch in den "Data Model Viewer"gewechselt werden, hier sind die einzelnen Datensätze in Tabellenform mit Verknüpfungen zwischen den einzelnen Spalten dargestellt Nach jeder Änderung im "Data Manager"müssen die Daten erneut wenige Sekunden geladen werden, bevor weiter mit ihnen gearbeitet werden kann.

Somit stellt Qlik eine ähnliche Bandbreite an Data Understanding Funktionen wie Python zur Verfügung, während die meisten Prozesse deutlich schneller und effizienter ablaufen. Es treten primär Probleme bei der Verarbeitung von JSON Datein und bei sehr großen Datein über einem Gigabyte auf.

2.3 Data Preperation

Im Data Preparation-Schritt geht es darum, die Daten für die Analyse vorzubereiten, indem sie gereinigt, transformiert und in das richtige Datenformat gebracht werden, um sicherzustellen, dass sie für den Modellierungsschritt geeignet sind.

2.4 Modelling

Im Modelling-Schritt werden die vorbereiteten Daten verwendet, um Visualisierungsmodelle zu erstellen. Ziel ist es, visuelle Darstellungen und Interaktionsmöglichkeiten zu entwickeln, die es ermöglichen, Muster, Trends und Beziehungen in den Daten zu erkennen und zu kommunizieren.

2.5 Sonstiges

Zusätzlich zu den drei Hauptphasen des Toolvergleichs gibt es einen Sonstiges-Schritt, der sich mit anderen wichtigen Aspekten wie Kosten, Benutzerfreundlichkeit, Support und anderen nicht direkt im CRISP-DM-Prozess enthaltenen Faktoren befasst.

3 SWOT Analyse ???!!!

4 Fazit

4.1 Python

4.2 Looker

4.3 Wann welches Tool

5 Quellen

5.1 Quellen Datensätze

1. Beispiel Webseite
2. Eine andere Beispiel Webseite

3. Eine andere Beispiel Webseite
4. Eine andere Beispiel Webseite
5. Eine andere Beispiel Webseite
6. Eine andere Beispiel Webseite
7. Eine andere Beispiel Webseite
8. Eine andere Beispiel Webseite
9. Eine andere Beispiel Webseite
10. Eine andere Beispiel Webseite

5.2 weitere Quellen

11. Einige der Fragen an denen der Vergleich durchgeführt wurde wurden von hier übernommen