Inhaltsverzeichnis

1. Einführung

Wie bei vielen unternehmensbezogenen Prozessen, geht es darum mit möglichst geringem Aufwand bestmögliche Ergebnisse liefern zu können.
Genauso ist es auch bei Datenwissenschaftlern. Man muss wissen, welche Tools sich am besten für die jeweiligen Anwendungszwecke eignen, um Daten adäquat visualisieren zu können.
Dementsprechend soll diese Ausarbeitung eine SWOT-Analyse der beiden Tools Python und Qlik beinhalten, damit veranschaulicht werden kann, in welchen Bereichen diese beiden Nutzmittel ihre Vor- und Nachteile haben und wie sie genutzt werden können.
Somit sollen zukünftige Entscheidungen für die Wahl des Visualisierungstools erleichtert werden, der erkenntlich werden soll, für welchen Einsatz das jeweilige Tool besser geeignet ist.

1.1. Python

Python ist eine universelle und höhere Programmiersprache, die den Anspruch hat, einen gut lesbaren, knappen Programmierstil zu fördern.
Für die Data Visualization wird sie zudem sehr häufig verwendet, da sie mit ihren diversen Bibliotheken eine umfassende Grundlage für das Data Modelling darstellt.
Für den ausgewählten Datensatz soll nun im Folgenden und unter Berücksichtigung von mannigfaltigen Kriterien, analysiert werden, wie gut Visualisierungen mit Hilfe der Programmiersprache erstellt werden können.

1.2. Qlik

Qlik ist ein US-amerikanisches Softwareunternehmen, welches verschiedene Produkte bezüglich des Umgangs mit Daten anbietet.
Zudem umfasst es auch eine cloudbasierte End-to-End Analytics-Plattform für Echtzeit-Datenintegration und -analyse. Diese ist laut der Website von Qlik auf die Nutzung in Unternehmen ausgerichtet, bietet allerdings auch eine kostenlose Test-Version, welche auch von privaten Nutzern für 30 Tage getestet werden kann.
Laut Hersteller verspricht das Tool „eine Data Fabric für moderne Datenarchitekturen und Analysen der nächsten Generation“, wie auch „hochgradig, interaktive, kontextbezogene Dashboards mit blitzschnellen Steuerelementen.“
Jene Test-Version soll im weiteren Verlauf des Projekts, äquivalent zu Python, unter der Beachtung von ausgewählten Kriterien beleuchtet werden.

1.3. Datensatz

Der bereits erwähnte und bearbeitete Datensatz stammt von der Quelle „UN Data“ und befasst sich mit dem Thema Human Development, anhand welchem der spätere Vergleich zwischen Python und dem benutzten Tool, Qlik, durchgeführt werden soll.
Dies hat zum einen den Vorteil, dass die Online-Seiten der „United Nations“ als Quellen vertrauenswürdig sind und zum anderen von offiziellen Seiten die präsentierten Daten regelmäßig aktualisiert werden.
Die Daten werden hierbei jeweils in einzelnen kleinen Datensätzen präsentiert und zu jedem Land über einen Zeitraum von über 60 Jahren für einzelne Kategorien aufgeführt.
Additiv werden noch weitere Datensätze der WHO für den Toolvergleich herangezogen.
Dementsprechend kann nicht von einem großen Datensatz, sondern mehreren kleinen Datensätzen gesprochen werden, die im weiteren Verlauf der Analyse benutzt werden sollen.

Obwohl der Datensatz mehr oder weniger nur Mittel zum Zweck für die Tool-Analyse ist, so wurde dieser dennoch bewusst gewählt. Immerhin können mit dem Thema Human Development verschiedenste Fragen zum Leben der Menschen auf der Welt beantwortet werden, um basierend hierauf auch Entscheidungsfindungen in Politik und Wirtschaft in die Wege zu leiten. Somit betrifft dieses Thema indirekt die gesamte Gesellschaft und stellt in Kombination mit dem Vergleich von Qlik und Python eine interessante Aufgabe dar.

2. Vergleich Python und Qlik

Um den Vergleich der beiden Tools objektiv durchführen zu können, wird hierfür ein Bewertungsbogen hinzugenommen, der vielfältige Kriterien für die Analyse bereithält.
Diese orientieren sich am CRISP-DM Prozess und werden in die einzelnen Teilschritten von diesem unterteilt.
Des Weiteren sollen auch spezifische Visualisierungen zum Thema Human Development erstellt werden, wofür einzelne Fragen aufgestellt werden.
Hierfür sollen für jeweils beide Tools die Resultate möglicher Visualisierungen gezeigt werden und welche Fragen weniger adäquat mit Visualisierungen basierend auf dem Datensatz beantwortet werden können.

2.1. Business Understanding

Zu diesem ersten Schritt des CRISP-DM Prozesses können keine Kriterien angewandt werden, da das Business Understanding, in diesem Falle also das Thema Human Development, nicht in den einzelnen Tools durchgeführt wird. Hierfür kann eine Internet-Recherche benutzt werden, allerdings eignen sich verschiedene Visualisierungen nicht dafür, um das untersuchte Thema an sich zu erklären.

2.2. Data Understanding

Anders als das Business Understanding, kann das Data Understanding als erster Schritt mit Kriterien für die beiden Tools dienen. Hierbei soll vor allem untersucht werden, inwiefern es möglich ist, neue Daten in die Tools hinzuzufügen, übersichtlich anzuordnen und diese für erste Visualisierungen ohne weiteren Aufwand zu benutzen.

In Qlik funktioniert das Hinzufügen von einzelnen Datensätzen über ein sogenanntes Datenmanager-Feld, in welchem beispielsweise einzelne CSV-Dateien manuell ausgewählt und in den jeweiligen Ordner geladen werden können. Dieser Schritt ist durch eine gute Beschreibung der verschiedenen Features gut durchzuführen, selbst wenn man zuvor noch nie mit der Oberfläche gearbeitet hat.
Des Weiteren kann hierbei festgehalten werden, dass sich neue Daten auch in sehr geringer Zeit einfügen lassen. Bei den für die Analyse benutzen CSV-Dateien, sind in diesem Prozess nur wenige Sekunden verstrichen.
In der Hauptübersicht, wo auch die verschiedenen Visualisierungstypen zu finden sind, werden dann die einzelnen Kategorien der geladenen Datensätze in der Seitenleiste angezeigt, wodurch ein erster Überblick über die vorhandenen Daten möglich ist.

Solange nur ein Datensatz mit wenigen Kategorien geladen wurde, ist dies auch recht übersichtlich. Jedoch muss bei größeren Datensätzen mit mehreren Spalten etwas genauer gesucht werden, um spezifische hiervon identifizieren zu können. Wurden allerdings mehrere Datensätze in den gleichen Ordner geladen, kann auch zwischen einzelnen Datensätzen gewählt werden, welche an der Seitenleiste angezeigt werden sollen, was zusätzliche Übersicht garantiert.
Insofern die geladenen Datensätze dann schon nutzbare Daten enthalten, können im Folgenden auch ohne geringen Aufwand erste Testvisualisierungen für die initiale Sichtung und Einschätzung über die Qualität der Daten getätigt werden.
In dieser Hinsicht hat Qlik in der Analyse mehrere Kriterien durchaus erfüllt.

Gegenüberstellend sollen die zu Beginn genannten Kriterien nun auch auf die Programmiersprache Python angewandt werden. Hierfür wird allerdings immer vorausgesetzt, dass die nötigen Programmierkenntnisse vorhanden sind.
Unter Berücksichtigung dessen funktioniert auch hier das Einfügen von neuen Files mit Hilfe der Bibliothek „pandas“ relativ einfach, wofür auch keine richtigen Datengrößen als Richtlinie dienen können, ab wann das Laden von Datensätzen nicht mehr möglich ist. Dies funktioniert äquivalent zu Qlik ohne übermäßige Ladezeiten. Hierbei beansprucht jedoch das Schreiben des notwendigen Codes für diesen Schritt einige Minuten mehr als das Auswählen von Dateien für den Datenmanager in Qlik.
Weiter ist in Python auch eine initiale Datenexploration durch die head()- und info()-Funktionen möglich, um einen schnellen Überblick zu erhalten und Null-Werte anzeigen zu lassen. Allerdings weist die Übersichtlichkeit der geladenen Daten hierbei Defizite auf, da vor allem bei umfangreicheren Datensätzen das Auffinden spezifischer Zeilen oder Spalten per bloßem Auge nicht direkt gewährleistet ist.
Nichtsdestotrotz bietet Python einen anderen Vorteil, der so nicht direkt in Qlik existiert: das Laden durch eine API, sodass nicht einzelne Files manuell hinzugefügt werden müssen.
Damit erfüllt auch Python einige der obigen Bewertungsmaßstäbe. Zwar erbringen das Schreiben des Codes für das Laden und die etwas fehlende Übersichtlichkeit Nachteile gegenüber Qlik, werden aber durch den möglichen Einsatz von APIs ausgeglichen.

2.3. Data Preparation

Im dritten Schritt des CRISP-DM Prozesses wird für gewöhnlich Bezug auf die Präparation der Daten genommen, sodass im weiteren Arbeiten mit dem jeweiligen Datensatz keine Probleme entstehen. Dabei muss vor allem Wert auf das Data Cleaning gelegt werden, ob und inwiefern Inkonsistenzen, Null-Werte, einzelne Zeilen oder Spalten entfernt werden können.

Aber auch das Prüfen, ob Datenanreicherungen, Strukturänderungen und Datentypänderungen am Datensatz vorgenommen werden können, darf hierbei nicht zu kurz geraten.

In Python ist das erwähnte Data Cleaning zum Beispiel recht flexibel. Ist die nötige Erfahrung im Umgang mit der Sprache vorhanden, können unterschiedliche Befehle benutzt werden, um gewünschte Werte zu entfernen.
So konnten die gewählten Datensätze zum Thema Human Development sehr schnell von Null-Werten und anderen unbrauchbaren Inhalten gereinigt werden, um zeitnah erste Visualisierungen erstellen zu können. Auch die Umstrukturierung der einzelnen Datenmengen war ohne Probleme möglich. Hierfür erwiesen sich primär Online-Foren und ChatGPT als sehr hilfreich und brauchbar.
Außerdem konnten zusätzlich gewünschte Kriterien als Spalten sehr gut zu den einzelnen Datensätzen hinzugefügt werden, um somit ein größeres Kategorienspektrum für die spätere Auswertung zu ermöglich und auch Datentypänderungen mit Python ohne Probleme ausgeführt werden.

Im Gegensatz hierzu erwies sich Qlik in diesem Schritt als weniger brauchbar, da innerhalb des Tools ein wirkliches „Cleanen“ von importierten Datensätzen gar nicht erst möglich war. Zwar konnten im Prozess des Visualisierens einige Kategorien oder Werte gefiltert, aber nicht gänzlich aus den Datensätzen gelöscht, werden.
Weiter waren auch Änderungen an Datentypen oder gar der Struktur der einzelnen Datensätze nicht durchzuführen, sodass die importierten Daten zuvor gänzlich mit Python gecleaned werden mussten, da dies im Tool nicht mehr möglich war.
Auch das Zusammenfügen einzelner Datenmengen erwies sich als eher unpraktisch. Zwar konnten diese im sogenannten Datenmanager verknüpft werden, allerdings traten dann Probleme bei der Modellierung auf, da die Spalten und Zeilen nicht genau übereinstimmten und wie bereits erwähnt, das Ändern dieser im Tool selbst nicht möglich ist.

Dementsprechend kann für diesen Prozessschritt festgehalten werden, dass Python für die Anpassung von Daten für die weitere Benutzung definitiv die bessere Alternative gegenüber Qlik darstellt.

2.4. Modelling

2.5. Skalierbarkeit