

TIQ.View

Interactive Data Profiling



TIQView 1.0

Benutzerhandbuch

Stand: 07.03.2008
Version 1.0

INHALTSVERZEICHNIS

1. EINFÜHRUNG	2
1.1. BEVOR SIE STARTEN	2
1.2. SYSTEMVORRAUSSETZUNGEN.....	3
1.3. KONTAKT	5
2. TIQVIEW.....	6
2.1. TIQVIEW STARTEN.....	6
2.2. AUSWAHL TABELLEN UND SPALTEN	9
2.3. AUSWERTUNG.....	12
2.3.1. <i>Aufbau der TIQView-Anwendungen.....</i>	<i>13</i>
2.3.2. <i>Filter.....</i>	<i>14</i>
2.3.3. <i>Aufheben von Filtern.....</i>	<i>17</i>
2.3.4. <i>Sperren und Entsperrn einer Auswahl.....</i>	<i>17</i>
2.3.5. <i>Erstellen und Löschen von Bookmarks.....</i>	<i>17</i>
2.3.6. <i>TIQView-Funktionen / Workflow</i>	<i>18</i>
3. ÜBUNGSBEISPIELE.....	29
3.1. ANALYSE FORMATVORKOMMEN VON POSTLEITZAHLEN FÜR KUNDEN AUS DEN USA	29
3.2. IDENTIFIZIERUNG VON MÖGLICHEN SCHLÜSSELATTRIBUTEN	31
3.3. WERTEVORKOMMEN	32
3.4. NULLWERT-ANALYSE.....	34

1. Einführung

1.1. Bevor Sie starten

TIQView - Interaktives Data Profiling an jedem Ort zu jeder Zeit!

Die TIQ Solutions GmbH hat mit *TIQView* eine spezielle Datenqualitäts-(DQ) Applikation entwickelt, mit der interaktives Data Profiling unabhängig von Ort und Zeit kein Problem mehr ist. Sie können ab sofort Ihre Daten mit auf die (Dienst-)Reise nehmen und ganz einfach auf Ihrem Laptop prüfen, wie es um die Qualität Ihrer Abteilungsdaten bestellt ist. Mit *TIQView* schaffen Sie sich schnell einen ersten Überblick, welche Qualitätsprobleme in Ihrer Datenwelt existieren und wo Sie mit gezielten Verbesserungsmaßnahmen ansetzen müssen. Dadurch werden auch ökonomische Effekte schnell wirksam, die wiederum die Akzeptanz solcher DQ-Aktivitäten für die Zukunft erhöhen!

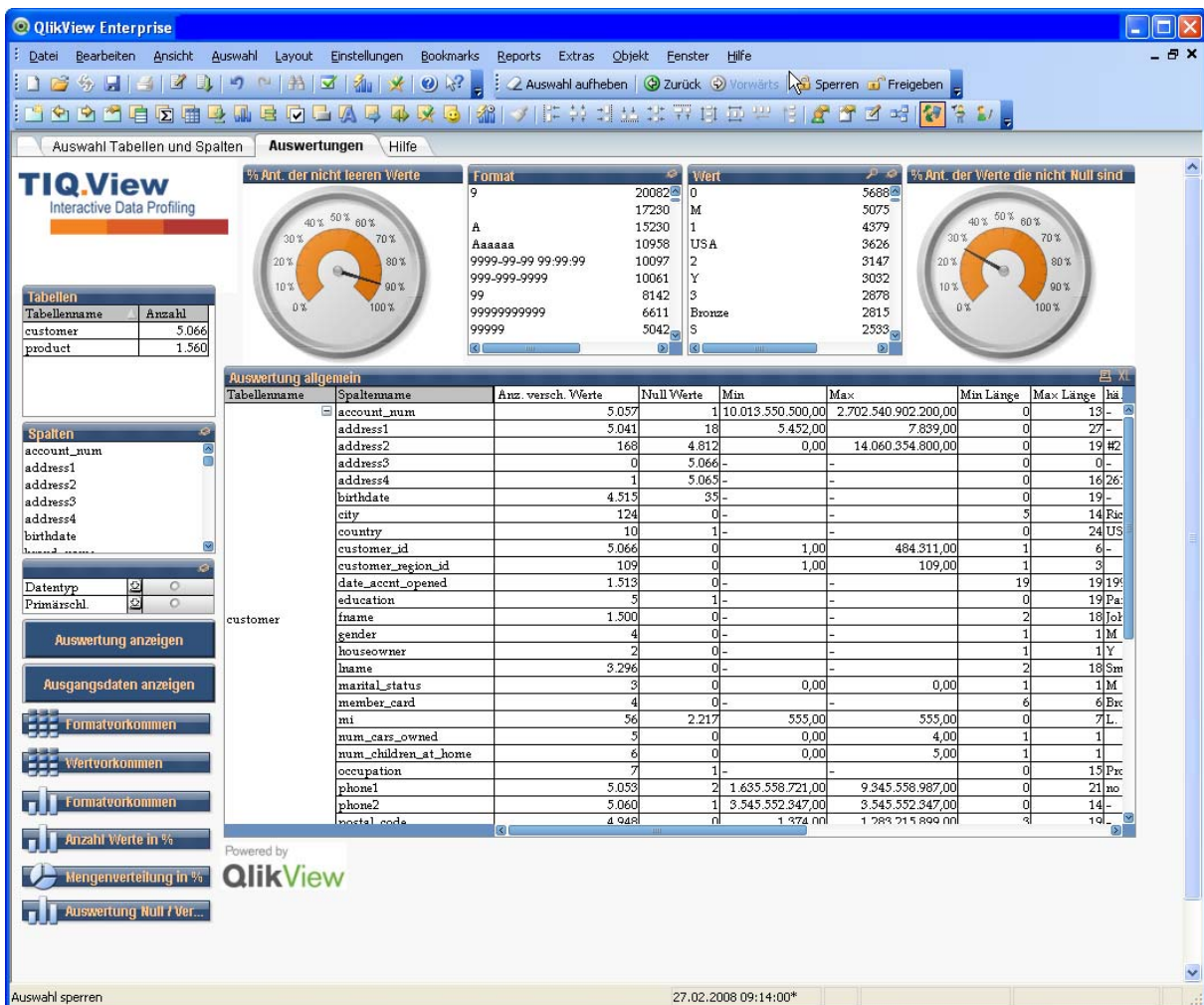


Abbildung 1: TIQView-Analyseoberfläche

TIQView ist ein Analysewerkzeug für Fachanwender unabhängig von bestimmten Branchen und Unternehmensbereichen, die sich mit Fragen der

Datenqualität auseinandersetzen. Die funktionalen Vorteile gegenüber kostenintensiven herkömmlichen Profiling-Werkzeugen sind:

- Intuitiver und übersichtlicher Zugang zu den wirklichen Datenqualitätsproblemen
- Durchführung von Analysen direkt in den Fachbereichen, unabhängig von der IT-Abteilung
- Einfach, flexibel und schnell erstellte aussagekräftige Auswertungen
- Kreative Erschließung von komplexen, tabellenübergreifenden Zusammenhängen
- Viele zusätzliche Informationen zu den selektierten Daten
- Unkomplizierte Kommunikation mit den Daten- und Prozessverantwortlichen
- Automatisches Report Update

TIQView bietet Ihnen dabei allgemeine DQ-Auswertungen wie z.B.:

- Identifizierung von Schlüsselattributen
- Werte- und Format-Analyse
- Datentypen-Analyse
- Nullwert-Analyse
- Häufigkeits-Analyse
- Feldlängen-Analyse
- Wertebereichs-Analyse
- Sonstige statistische Verfahren
- Drill Down zu den jeweils relevanten Datensätzen

TIQView kann wesentlich mehr als reine Spalteninhaltsanalysen Ihrer Datenbank-Tabellen! Durch die funktionale Flexibilität gelingt es in nur wenigen Minuten, komplizierte datenqualitätsspezifische Zusammenhänge zwischen Spalten verschiedener Datenbanktabellen herzustellen. Zudem finden Sie damit unterschiedliche Möglichkeiten, sich den Problemen analytisch zu nähern, so dass Sie immer wieder neue Perspektiven bzw. Fragestellungen aufwerfen und vertiefen können. Durch die übersichtliche Gestaltung der einzelnen Funktionen, Selektionen und Daten in der graphischen Oberfläche hat der Anwender auch bei Detailanalysen immer den gesamten Kontext der Auswertungen im Blick.

Diese permanente Übersicht ermöglicht einen schnellen, intuitiven und kreativen Umgang mit TIQView. Bei Bedarf bieten wir Ihnen eine Anbindung an ein Koordinationswerkzeug, mit dem Sie auf sehr effiziente Art und Weise die aus den recherchierten DQ-Problemen abgeleiteten Maßnahmen definieren, verwalten und überwachen können.

1.2. Systemvoraussetzungen

TIQView basiert auf der Business Intelligence Software Qlikview von der Firma

Qliktech. Für den Betrieb von TIQView wird die Anwendung Qlikview in der Version 7 oder 8 mit dem Produktlevel Analyzer auf Ihrem System benötigt. Ein Qlikview-Installationsprogramm ist in der Installation von TIQView enthalten. Beim Erwerb von TIQView erhalten Sie eine Lizenz des Qlikview Analyzers, welche ausschließlich für TIQView genutzt werden darf.

Sollten sie noch keine TIQView- oder Qlikview-Lizenz haben, können Sie eine 15 tägige Testversion zur Evaluierung von TIQView nutzen. Registrieren Sie sich dafür unter www.tiqview.de.

Um Ihre Daten in TIQView zu laden, muss ein an Ihre Datenquellen angepasstes CSV erzeugt werden. Der dafür notwendige TIQView-Skriptgenerator setzt eine Java-Laufzeitumgebung voraus, welche ebenfalls im Lieferumfang von TIQView enthalten ist.

Folgende Voraussetzungen sind für Nutzung von TIQView notwendig:

Hardware:

Ihr Computer sollte folgendermaßen ausgestattet sein:

- Mindestens Pentium II Prozessor (Empfehlung: Pentium 4 oder besser)
- QlikView nutzt die Vorteile moderner Multi-Core-Prozessoren vollständig aus, daher empfiehlt Qliktech diese Prozessoren
- Grafikkarte für GA-Auflösung in Farbe oder besser
- Monitore mit einer Auflösung von mindestens 1024 x 768 Punkten
- Maus oder äquivalentes Windows-gestütztes Pointer System
- CD-ROM-Laufwerk (falls Sie von CD installieren möchten)
- Festplatte mit mindestens 150 MB freiem Speicher
- Mindestens 512 MB RAM für die 32-Bit-Versionen von Windows
- 2 GB RAM für die 64-Bit-Versionen von Windows
- Der benötigte Speicher ist nahezu proportional zur gewünschten Datenmenge

Betriebssystem:

Sie benötigen eines der folgenden Betriebssysteme:

- Microsoft Windows 2000 TM
- Microsoft Windows 2000 Server TM
- Microsoft Windows 2003 Server TM
- Microsoft Windows XP TM
- Microsoft Windows Vista TM

1.3. Kontakt

Haben Sie Fragen oder Anregungen zu TIQView oder zum Thema Datenqualität?

Kontaktieren Sie uns: support@tiq-solutions.de

2. TIQView

2.1. TIQView starten

Die TIQView Anwendung steht in verschiedenen an die Bildschirmauflösungen optimierten Versionen zur Verfügung. Im TIQView-Programmverzeichnis (z.B. C:\Programme\TIQView) befinden sich die Programmdateien mit **TIQView 1024 x 768.qvw** angepasst an die Bildschirmauflösung 1024 x 768 Pixel sowie die Programmdatei **TIQView 1280 x 1024.qvw** für Bildschirmauflösungen ab 1280 x 1024 Pixel und höher.

Starten Sie TIQView durch Doppelklick auf die Ihrer Bildschirmauflösung entsprechenden TIQView Programmdatei im TIQView-Programmverzeichnis (z.B. C:\Programme\TIQView).

Alternativ können Sie TIQView auch innerhalb von QlikView öffnen. Starten Sie dazu Qlikview. Gehen Sie in das Menü „**Datei** → **Öffnen**“ und wählen Sie aus Ihrem TIQView-Verzeichnis die Datei **TIQView 1024 x 768.qvw** bzw. **TIQView 1280 x 1024.qvw** aus. Geben Sie als Benutzername *user* ein (die Eingabe eines Passwortes ist nicht notwendig).

Sie gelangen auf die Startseite:

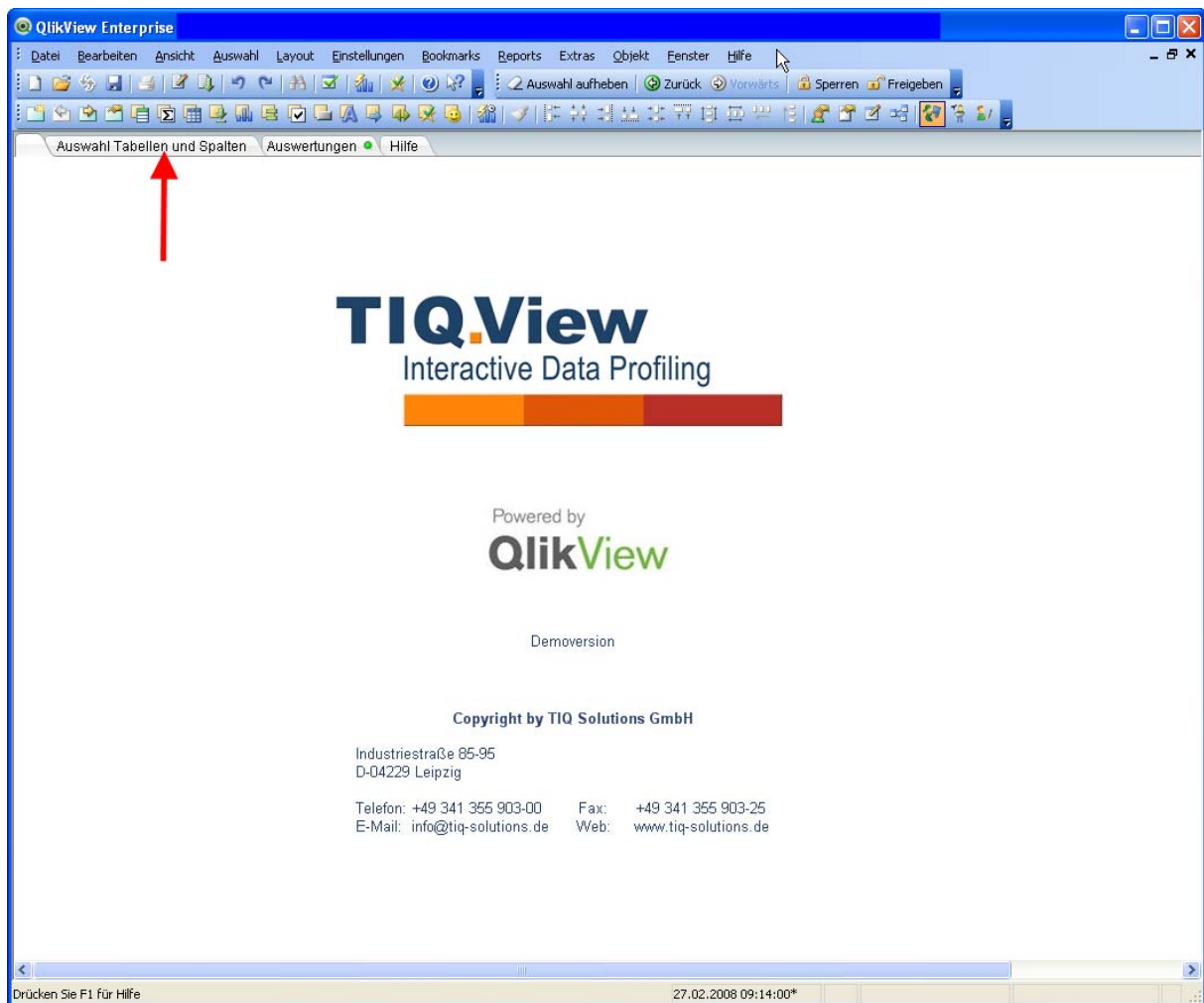


Abbildung 2: TIQView starten

Durch klicken auf den Registerreiter „**Auswahl Tabellen und Spalten**“ kommen Sie auf die Seite für die Auswahl der Tabellen und Spalten:

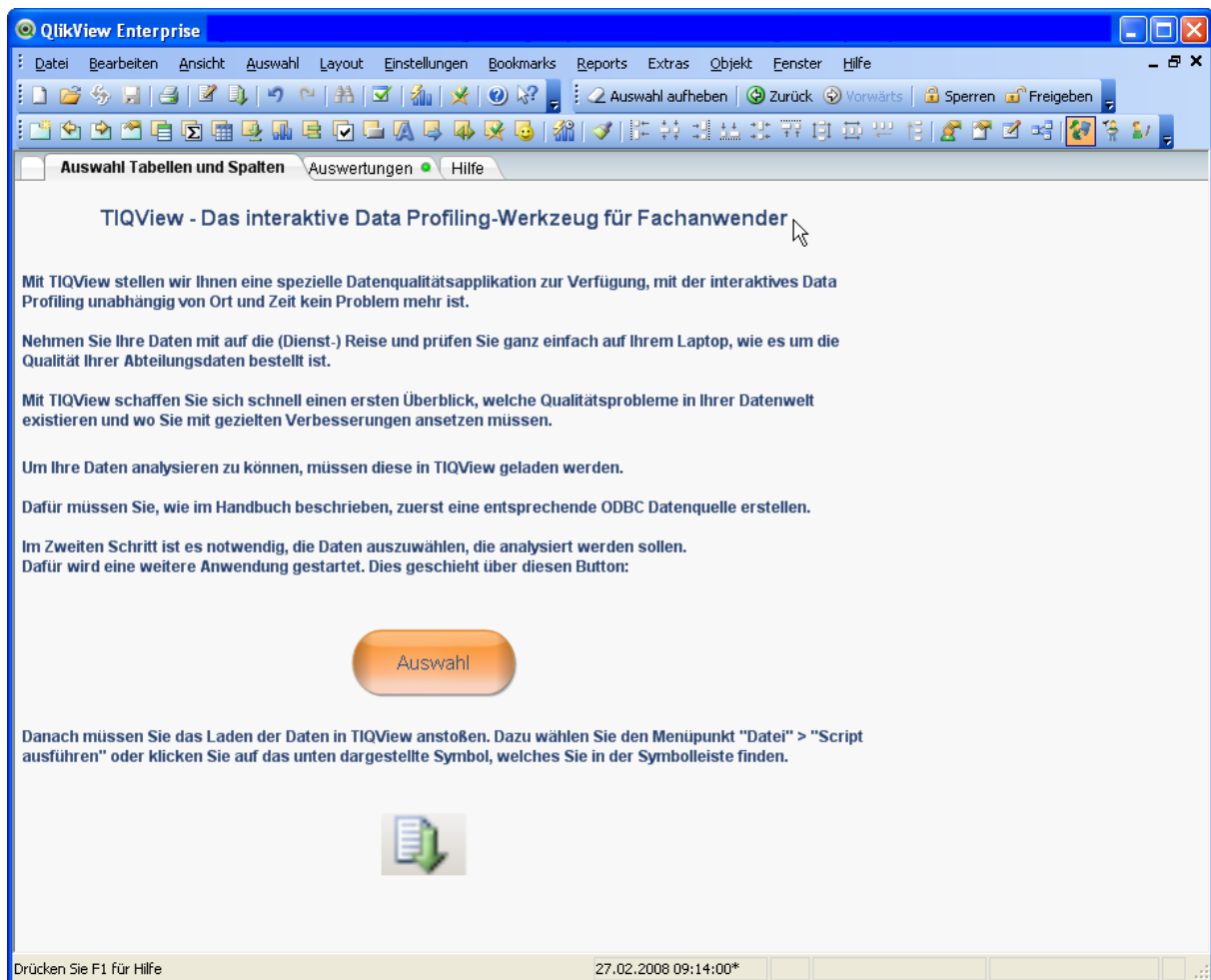


Abbildung 3: Auswahl Tabellen und Spalten

Durch klicken auf den oben dargestellten Button wird die Auswahl der Tabellen und Spalten aus der Datenquelle aktiviert.

2.2. Auswahl Tabellen und Spalten

Dieser Abschnitt hat für die Demoversion keine Relevanz!

Nach dem Starten des TIQView-Composers finden Sie die folgende Programmoberfläche vor sich:

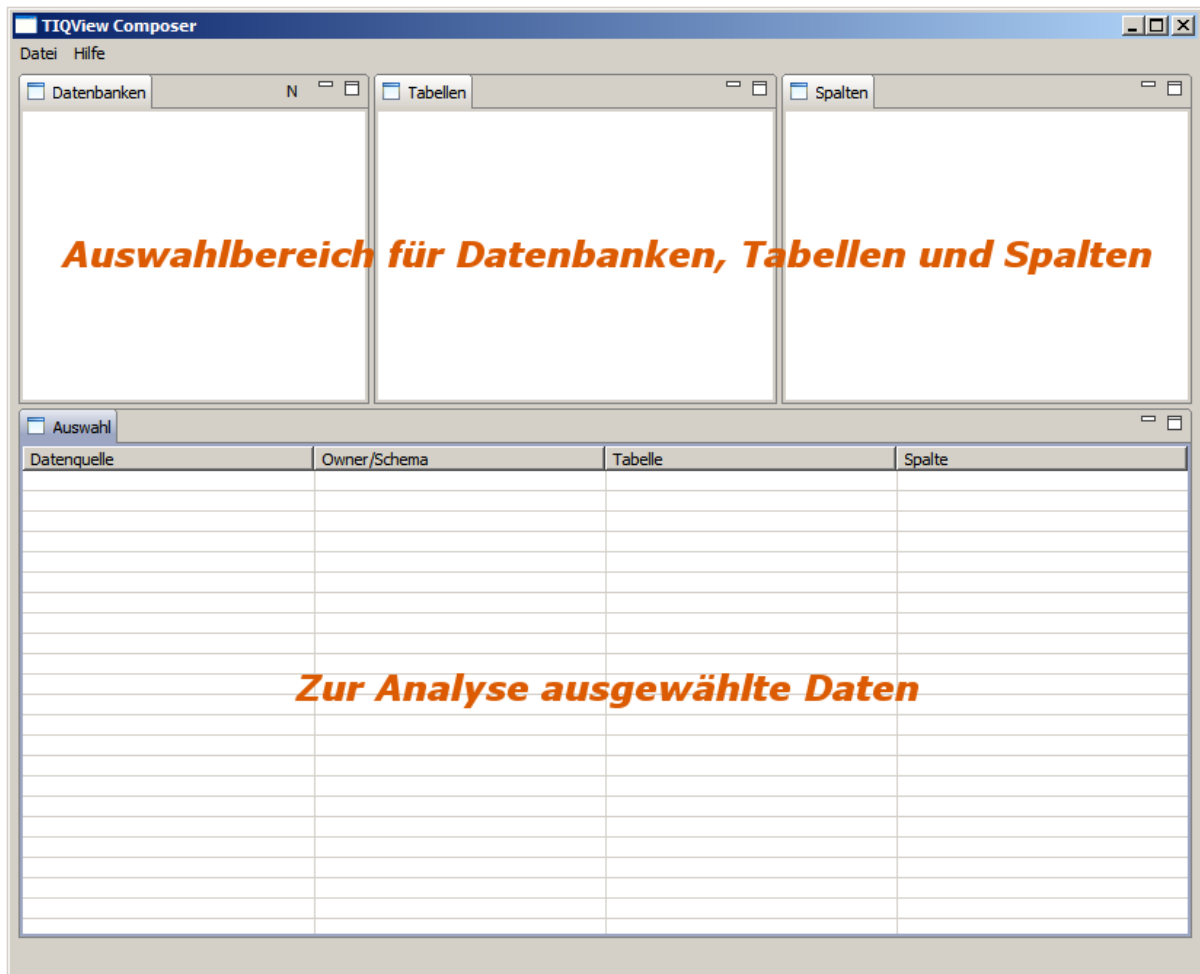


Abbildung 4: Standard-Arbeitsoberfläche des TIQView-Composers

Um eine neue Datenbankverbindung herzustellen, wählen Sie im Menü „**Da-tei**“ den Unterpunkt „**Neue Datenbankverbindung**“, dadurch öffnet sich der nachfolgend dargestellte Dialog:

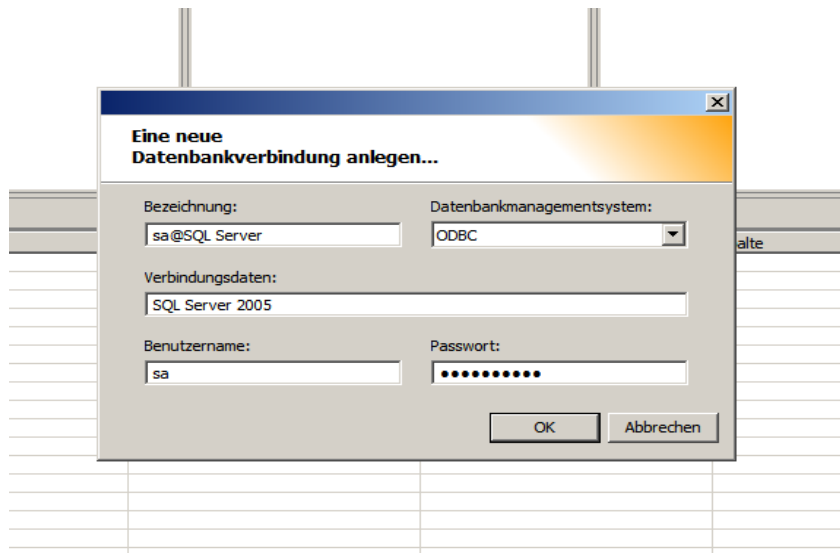


Abbildung 5: Anlegen einer neuen Datenbankverbindung

Um eine neue Verbindung zu einer vorhandenen ODBC-Quelle aufzubauen, gehen Sie wie folgt vor:

1. Geben Sie im Feld „**Bezeichnung**“ der Abbildung 5 eine beliebige für Sie nachvollziehbare Bezeichnung der Datenbankverbindung an.
2. Wählen Sie „**ODBC**“ im Feld Datenbankmanagementsystem aus.
3. Geben Sie im Feld Verbindungsdaten die Systembezeichnung der ODBC Quelle ein. Diese Bezeichnung finden Sie in Windows unter „**Sys-temsteuerung** → **Verwaltung** → **Datenquellen (ODBC)**“.

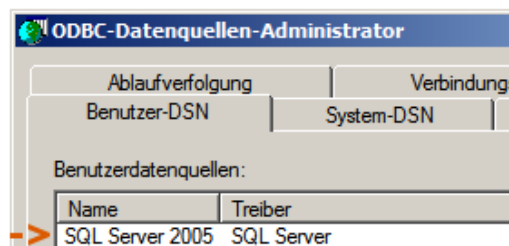


Abbildung 6: ODBC-Quellen unter Windows

4. Geben Sie in den Feldern „Benutzername“ und „Passwort“ der Abbildung 5 Ihre Verbindungsinformationen ein. Bei einigen ODBC-Quellen (z.B. Excel) sind diese Angaben nicht notwendig.
5. Klicken Sie auf „**OK**“.

Wenn Sie nun auf die neu erstellte Verbindung zur ODBC-Quelle klicken (siehe Abb. 11) werden die für Sie verfügbaren Schema/Owner eingelesen. Über

diese können Sie zu den dazugehörigen Tabellen und Spalten navigieren.

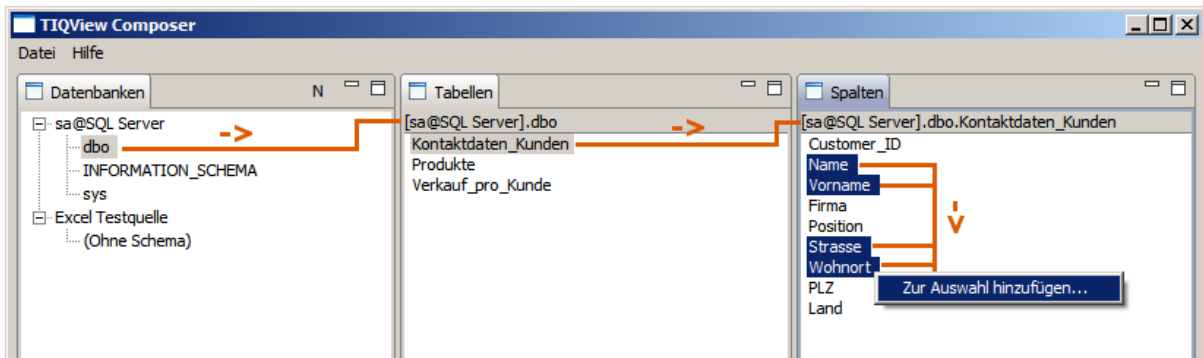


Abbildung 7: Auswahl und Navigation im Explorerbereich des TIQView-Composers

Um mehrere Spalten auszuwählen, halten Sie dabei bitte die Taste **[STRG]** gedrückt. Haben Sie die Spalten ausgewählt, können Sie mit einem Klick auf die rechte Maustaste im Popup-Menü den Punkt „**Zur Auswahl hinzufügen**“ auswählen und damit die gewählten Spalten in das Auswahlfenster einfügen.

Im Fenster „**Auswahl**“ unterhalb des Explorerbereichs im TIQView-Composer finden Sie eine Zusammenstellung aller bis dahin gewählten Spalten, geordnet nach Datenquelle, Owner/Schema und Tabelle. Mit einem Klick auf die rechte Maustaste innerhalb des Auswahlfensters können Sie ihre Auswahl über das Popup-Menü mit „**Auswahl speichern...**“ speichern.

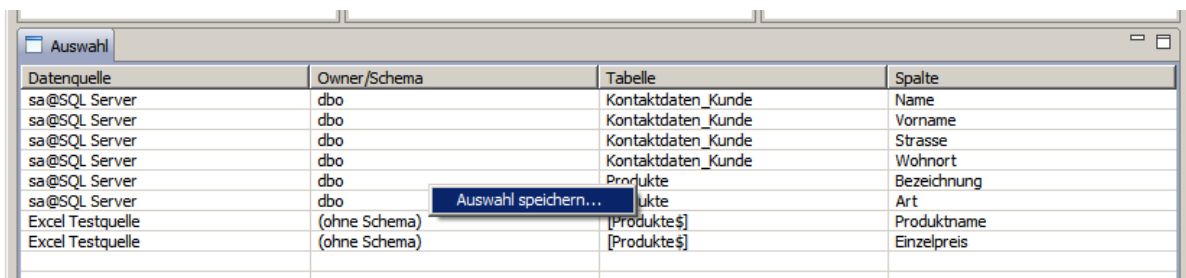


Abbildung 8: Zusammenstellung der ausgewählten Datenfelder für die TIQView-Analyse

An diesem Punkt können Sie den TIQView-Composer schließen oder in den Hintergrund verschieben und das Laden der Daten in TIQView über den Menüpunkt „**Datei** → **Script ausführen**“ sowie in TIQView selber mit der Auswertung beginnen.

2.3. Auswertung

Von der Seite „Auswahl Tabellen und Spalten“ gelangen Sie durch Klick auf den Registerreiter „Auswertungen“ direkt zur Analyseoberfläche.



Abbildung 9: Analyseoberfläche von TIQView

2.3.1. Aufbau der TIQView-Anwendungen

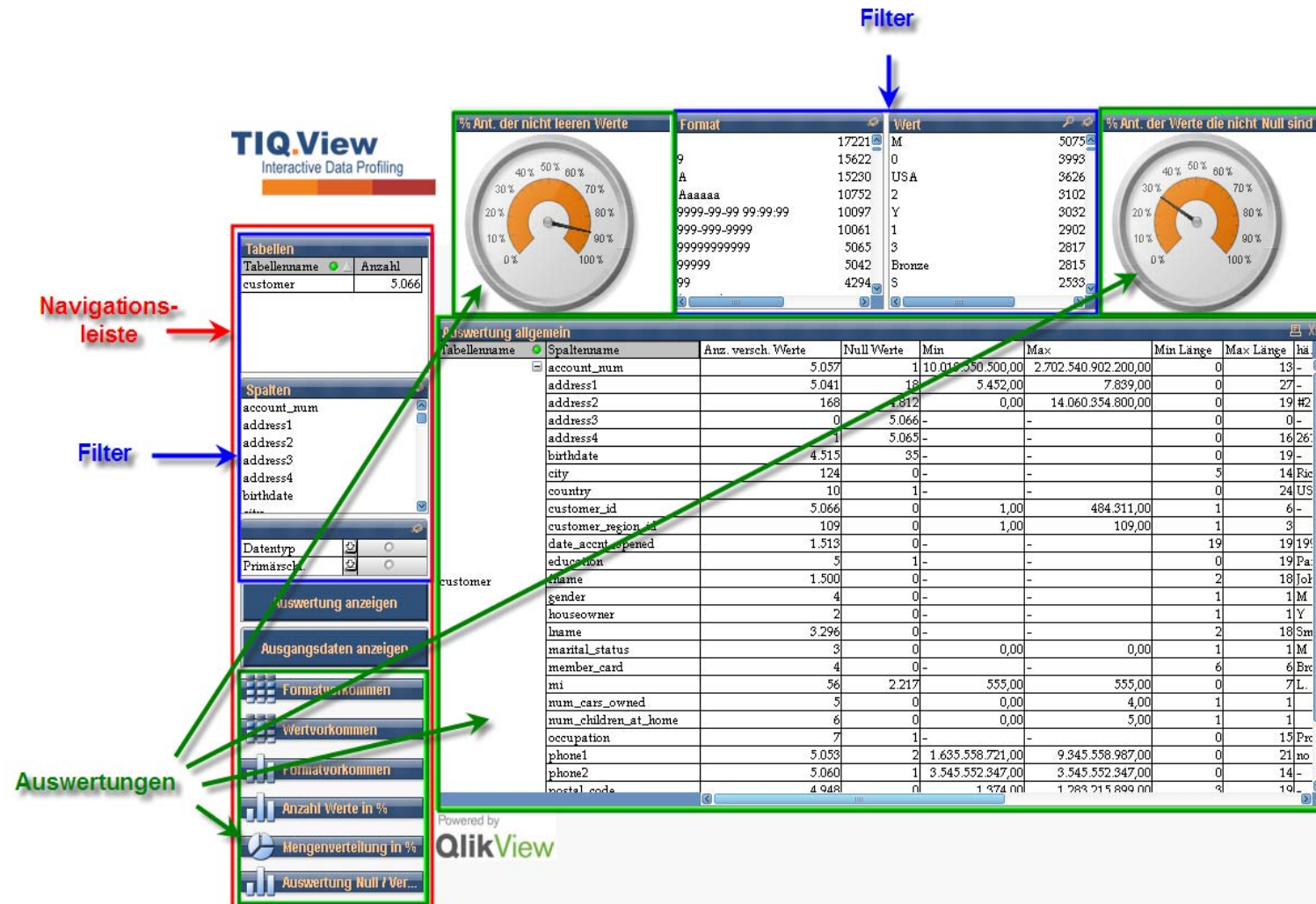


Abbildung 10: Überblick TIQView-Cockpit

Die Oberfläche der TIQView-Anwendung ist in zwei Bereiche aufgeteilt. Auf der linken Seite befindet sich die Navigationsleiste mit Filtermöglichkeiten und minimierten Auswertungen, die per Klick auf den jeweiligen Button aufgerufen werden können. Daneben wird der Auswertungsbereich dargestellt, der je nach Auswahl verschiedene Analysen enthalten kann.

2.3.2. Filter

In TIQView haben Sie die Möglichkeit, Einschränkungen der Daten in so genannten

Listboxen

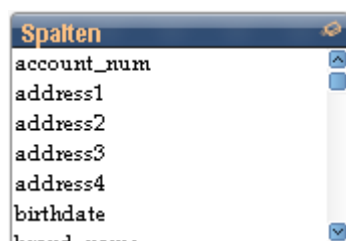


Abbildung 11: Listbox

oder

Multiboxen

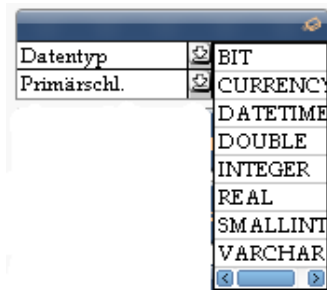


Abbildung 12: Multibox

und in den Tabellen oder Diagrammen der Auswertungen selbst vorzunehmen.

TIQView besitzt darüber hinaus eine intuitive Bedienoberfläche: Ein Klick auf die auszuwählenden Werte in Listboxen, Multiboxen und/oder den verschiedenen Auswertungen schränkt die angezeigten Werte auf die ausgewählten ein. Die Auswahl kann übergreifend in mehreren Auswahlboxen, Tabellen etc. erfolgen.

Bei der Auswahl sollte berücksichtigt werden, dass:

- eine Selektion von Werten grün hinterlegt wird
- Werte, welche passend zu bisherigen Selektionen wählbar sind, mit einem weißen Hintergrund versehen sind
- Werte, welche nach einer Auswahl nicht mehr gewählt werden können, grau hinterlegt sind, da sie mit der momentanen Auswahl nicht kombinierbar sind.

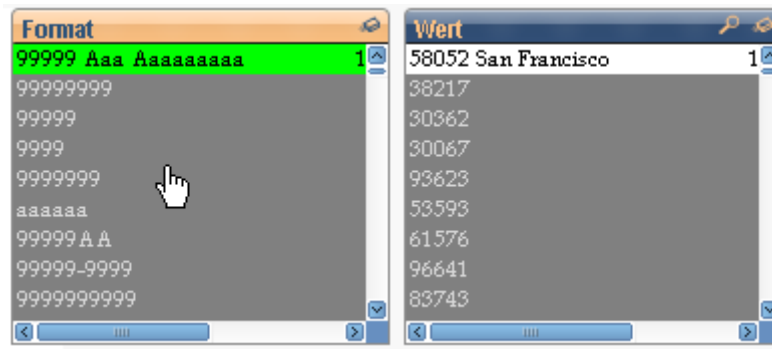


Abbildung 13: Auswahl in einer Listbox

- Werden Auswahlen in einer Tabelle getroffen, verschwinden alle Werte bis auf die selektierten. An einem kleinen grünen Punkt neben dem Spaltennamen sehen Sie, dass eine Selektion vorgenommen wurde

Auswertung allgemein		
Tabellenname	Spaltenname	Anz. versch. Werte
customer	account_num	5.057
	address1	5.041
	address2	168
	address3	0
	address4	1
	birthdate	4.515
	city	124
	country	10
	customer_id	5.066
	customer_region_id	109
	date_accnt_opened	1.513
	education	5
	fname	1.500
	gender	4

Abbildung 14: Auswahl in einer Tabelle

- Selektionen innerhalb einer Listbox muss man als „ODER“-Verknüpfung betrachten. In der unten dargestellten Abbildung wurden alle Datensätze ausgewählt die entweder den Wert „USA“ oder „Bronze“ enthalten

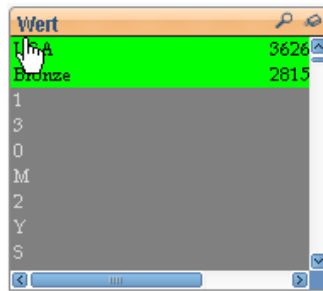


Abbildung 15: Selektion als "ODER"-Verknüpfung

- Selektionen in zwei oder mehreren Listboxen muss man als „UND“-Verknüpfung betrachten. In der unten dargestellten Abbildung wurden alle Datensätze ausgewählt die entweder das Format „aa aaa“, „aaa-aaa“ oder „999-999-9999a“ enthalten *und* in den Spalten „phone1“ oder „phone2“ vorkommen

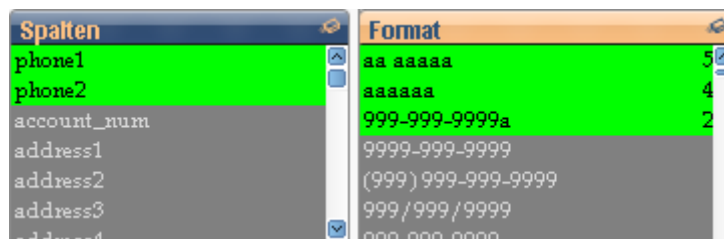


Abbildung 16: Selektion als "UND"-Verknüpfung

Sie können jeweils einen oder mehrere Werte auswählen. Bei mehreren Werten müssen Sie die **[STRG-Taste]** gedrückt halten oder einfach mit der gedrückten linken Maustaste über die zu selektierenden Werte ziehen. In Listboxen und Multiboxen kann nach speziellen Werten gesucht werden. Klicken Sie dafür auf die Titelleiste der Listbox oder auf den Dropdown-Pfeil einer Spalte in einer Multibox und beginnen Sie den Begriff zu tippen, nach dem Sie suchen. Es öffnet sich ein Fenster in dem Sie den Suchbegriff sehen. Alternativ können sie auch die Taste F3 drücken. Die Suche kann durch die Platzhalter „*“ (für mehrere Buchstaben) und „?“ (für einen Buchstaben) erweitert werden.

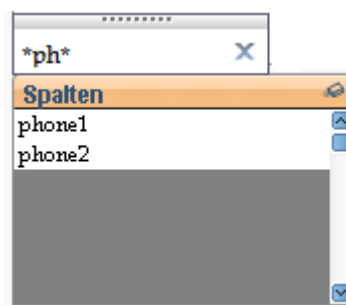


Abbildung 17: Suchen in Listboxen



Durch den Klick mit der **„rechten Maustaste auf eine Listbox“** können zudem folgende Aktionen auf die Felder der Listbox durchgeführt werden:

- **„Wählbare Werte auswählen“**
- **„Ausgeschlossene Werte auswählen“**
- **„Auswahl sperren“**: Sperrt die Auswahl in der Listbox, die gesperrten Werte können nicht durch **„Auswahl aufheben“** aufgehoben, sondern müssen durch **„Freigeben“** entsperrt werden
- **„Auswahl aufheben“**
- **„Auswahl in anderen Feldern auswählen“**, hebt Auswahl in anderen Auswahlboxen auf



Multiboxen sind komprimierte Darstellungen von Listboxen und werden wie diese behandelt.

2.3.3. Aufheben von Filtern

Es gibt verschiedene Möglichkeiten eine getroffene Auswahl aufzuheben:

- die in einer List- oder Multibox selektierten Werte können einzeln durch einen Klick auf das Radiergummi-Symbol in der Titelleiste der Listbox aufgehoben werden 
- ebenfalls ist es möglich die Auswahl eines Wertes durch nochmaliges Anklicken aufzuheben
- die gesamte Auswahl kann über den Button  in der Menüleiste von Qlikview aufgehoben werden

2.3.4. Sperren und Entsperrn einer Auswahl


Um zu verhindern, dass eine Auswahl durch **„Auswahl aufheben“** aufgehoben wird, kann eine Auswahl über  in der Symbolleiste gesperrt werden. Die so gesperrte Auswahl kann nur über  wieder freigeben und die Selektion danach aufgehoben werden.

Sie können die Auswahl in einzelnen List- oder Multiboxen oder auch die gesamte Auswahl sperren und freigeben. Bei einzelnen Sperrungen oder Freigeben gehen Sie bitte über das Menü der rechten Maustaste.

2.3.5. Erstellen und Löschen von Bookmarks

Um eine Auswahl dauerhaft zu speichern, können Sie Bookmarks erstellen. Diese Bookmarks können zu einem späteren Zeitpunkt wieder aufgerufen werden und die im Bookmark gespeicherten Selektionen werden angewen-

det.

Die Verwaltung von Bookmarks steuern Sie über den Menüpunkt Bookmarks. Zum Anlegen eines neuen Bookmarks können Sie über das Menü „**Bookmarks** → **neues Bookmark auswählen**“, in der Symbolleiste den Button  klicken oder sie nutzen die Tastenkombination **[STRG-B]**.

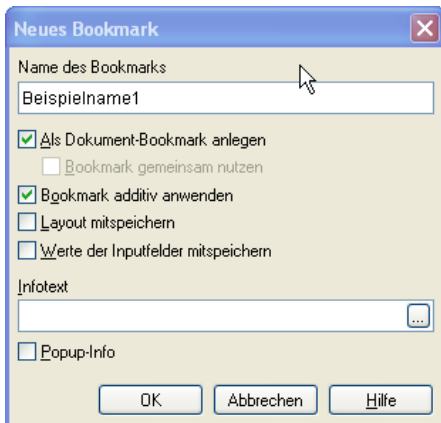


Abbildung 18: Anlegen eines Bookmarks

mark“ sind für die Verwendung von TIQView nicht relevant.

Vergeben Sie einen sinnvollen Namen für Ihr Bookmark. Um das Bookmark dauerhaft im Dokument zu speichern, setzen Sie den Haken bei „**Als Dokument-Bookmark anlegen**“.

Möchten Sie, dass die Auswahl des Bookmarks zu einer bestehenden Auswahl hinzugefügt wird, markieren Sie „**Bookmark additiv anwenden**“. Ist dieses Merkmal nicht aktiviert, wird eine bestehende Auswahl bei Aufruf des Bookmarks überschrieben. Weitere Optionen beim Dialog „**Neues Bookmark**“

Zum Aufrufen eines bestehenden Bookmarks wählen Sie unter dem Menüpunkt „**Bookmarks**“ das entsprechende Bookmark aus.

Bookmarks können ebenso unter dem Menüpunkt „**Bookmarks** → **Bookmarks löschen**“ gelöscht werden.

2.3.6. TIQView-Funktionen / Workflow

Tabellen

Tabellen	
Tabellenname	Anzahl
customer	5.066
product	1.560

Abbildung 19: Tabellenauswahl

Selektieren sie hier die Tabellen, die sie analysieren wollen. Erfolgt keine Auswahl werden alle aufgelisteten Tabellen betrachtet. Die Anzahl der Datensätze je Tabelle wird als zusätzliche Information angezeigt.

Spalten

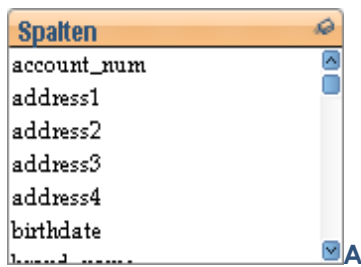


Abbildung 20: Spaltenaus-

Hier können Sie die Spalten auswählen, die Sie überprüfen möchten. Wurden keine Tabellen in der Tabellenbox ausgewählt, enthält die Listbox alle Spalten aller Tabellen, sonst stehen nur die Spalten der selektierten Tabellen zur Auswahl.

Datentyp und Primärschlüssel



Abbildung 21 Datentypen und Primärschlüssel

Möchten Sie nur bestimmte Datentypen betrachten (z.B. VARCHAR), wählen sie über das Dropdown Menü der Multibox den oder die entsprechenden Datentypen aus.

Innerhalb der oben genannten Multibox werden die Primärschlüssel der von Ihnen selektierten Datensätze angezeigt. Dies ist allerdings nur möglich, wenn entsprechende Primärschlüsselinformationen in der Datenquelle vorhanden sind. Ansonsten finden Sie dort einen künstlich erzeugten Primärschlüssel.

Auswertungen

Auswertung allgemein

Für die selektierten Daten werden in der Tabelle „Auswertung allgemein“ verschiedene statistische Aussagen pro Tabelle und Spalte getroffen:

- Anzahl verschiedener Werte
- Anzahl der Null-Werte
- die minimalen und maximalen Werte
- die minimalen und maximalen Feldlängen
- der am häufigsten auftretende Wert
- das am häufigsten auftretende Format
- die Anzahl numerischer und alphanumerischer Werte
- für numerische Werte: Standardabweichung, Median, Mittelwert

Auswertung allgemein										
Tabellenname	Spaltenname	Anz. versch. Werte	Null Werte	Min	Max	Min Länge	Max Länge	häufigster Wert	häufigstes Format	Anz
customer	account_num	5.057	1	10.013.550.500,00	2.702.540.902.200,00	0	13	-	9999999999	
	address1	5.057	2	5.452,00	7.839,00	0	27	-	9999 aaaaaa aaaaa	
	address2	168	4.812	0,00	14.060.354.800,00	0	19	#2		
	address3	0	5.066	-	-	0	0	-		
	address4	1	5.065	-	-	0	16	261 La Orinda Pl		
	birthdate	4.544	0	-102.434,00	368.990,00	4	7	-	9999-99-99 99:99:99	
	city	121	0	-	-	5	14	Richmond	aaaaaaa	
	country	13	1	-	-	0	24	USA	aaa	
	customer_id	5.066	0	1,00	484.311,00	1	6	-	9999	
	customer_region_id	109	0	1,00	109,00	1	3	-	72	99
	date_acct_opened	1.513	0	-2.658,00	34.695,00	5	5	33433	9999-99-99 99:99:99	
	education	5	1	-	-	0	19	Partial High School	aaaaaaa aaaa aaaaaa	
	fname	1.500	0	-	-	2	18	John	aaaaa	
	gender	4	0	-	-	1	1	M	a	
	houseowner	2	0	-	-	1	1	Y	a	
	lname	3.296	0	-	-	2	18	Smith	aaaaaa	
	marital_status	3	0	0,00	0,00	1	1	M	a	
	member_card	4	0	-	-	6	6	Bronze	aaaaa	
	mi	53	2.218	555,00	555,00	0	7	L	a.	
	num_cars_owned	5	0	0,00	4,00	1	1	2		9
	num_children_at_home	6	0	0,00	5,00	1	1	0		9
	occupation	7	1	-	-	0	15	Professional	aaaaaaaaaaaa	
	phone1	5.060	1	-	-	0	17	-	999-999-9999	
	phone2	5.060	1	3.545.552.347,00	3.545.552.347,00	0	14	-	999-999-9999	
	postal_code	4.948	0	1.374,00	1.283.215.899,00	3	19	-		99999
	state_province	15	1	-	-	0	10	CA	aa	
total_children	7	0	0,00	100,00	1	3		3	9	
yearly_income	8	0	-	-	7	13	\$30K - \$50K	\$99a - \$99a		
employee	birth_date	52	0	4.574,00	29.029,00	4	5	5147	9999-99-99 99:99:99	
	department_id	12	0	1,00	19,00	1	2	16		99
	education_level	5	0	-	-	15	19	Partial College	aaaaaaaaaaaa	
	employee_id	1.155	0	1,00	1.156,00	1	4	-		999
	end_date	0	1.155	-	-	0	0	-		

Abbildung 22: Allgemeine Auswertungstabelle

Formatvorkommen

Zum Aufruf der Auswertung Formatvorkommen doppelklicken Sie links in der Navigationsleiste auf „**Formatvorkommen**“.

Die Auswertung „Formatvorkommen“ zeigt sowohl in Diagrammform als auch als Tabelle, welche Formate in welcher Häufigkeit in den selektieren Tabellen und Spalten vorkommen. Die Formatvorkommen sind nach Häufigkeit sortiert. Natürlich können Sie die Sortierung jederzeit verändern, indem Sie auf die

entsprechende Spaltenüberschrift einen Doppelklick ausführen. Sie sehen die Sortierung anhand eines kleinen Pfeils, der in der Spaltenüberschrift erscheint. Durch diese Darstellung können Sie auffällige Formate leichter identifizieren.

Formatvorkommen				
Tabellenname	Spaltenname	Format	Anzahl	%
			5.000	100,00%
customer	address2	#A999	1	0,02%
customer	address2	9 AAAAAAAAAA	1	0,02%
customer	address2	999 Aaaaa Aaa	1	0,02%
customer	address2	9999 Aaaaaa Aaaa	1	0,02%
customer	address2	9999 Aaaaaaaaaa Aa	1	0,02%
customer	address2	A	1	0,02%
customer	address2	AAA Aaa #999	1	0,02%

Abbildung 24: Formatvorkommen

- Sie interessieren sich für einige auffällige Formatvorkommen? Markieren Sie die näher zu betrachtenden Formate einfach im Diagramm oder in der Tabelle und zoomen Sie somit in die Auswahl. Bedenken Sie, dass Sie für weitere Auswertungen die Auswahl wieder aufheben sollten.
- Zur Betrachtung der Daten in denen die auffälligen Formate vorkommen, klicken Sie auf den Button „**Ausgangsdaten anzeigen**“ und kontrollieren Sie den Datenbestand.

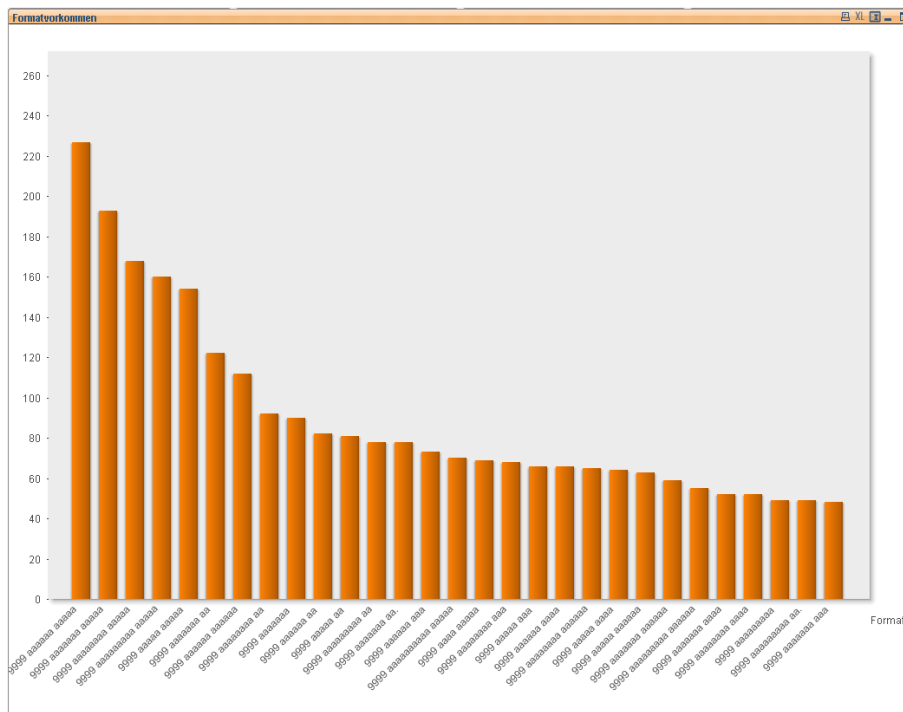
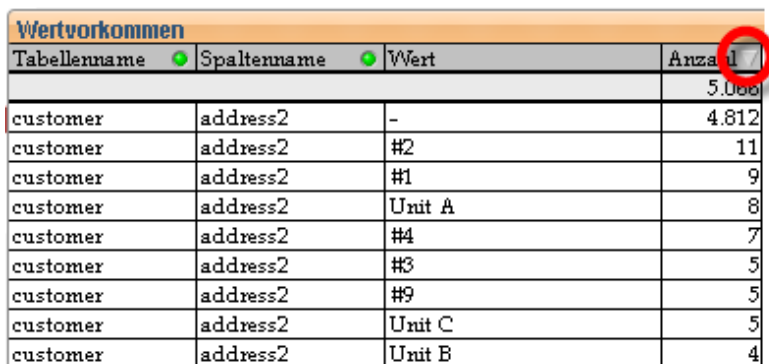


Abbildung 25: Formatvorkommen graphisch

Wertevorkommen

Zum Aufruf der Auswertung „Wertevorkommen“ doppelklicken Sie in der Navigationsleiste auf „**Wertevorkommen**“.

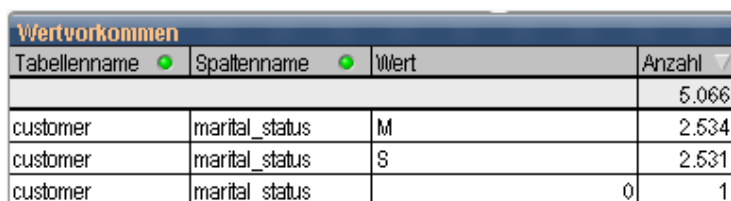


Wertevorkommen			
Tabellenname	Spaltenname	Wert	Anzahl
			5.066
customer	address2	-	4.812
customer	address2	#2	11
customer	address2	#1	9
customer	address2	Unit A	8
customer	address2	#4	7
customer	address2	#3	5
customer	address2	#9	5
customer	address2	Unit C	5
customer	address2	Unit B	4

Abbildung 26: Wertevorkommen 1

Die Auswertung „Wertevorkommen“ zeigt tabellarisch, welche Werte in welcher Häufigkeit in den selektierten Tabellen und Spalten vorhanden sind. Die Werte sind nach Häufigkeit sortiert. Natürlich können Sie die Sortierung jederzeit verändern, indem Sie auf die Spaltenüberschrift einen Doppelklick ausführen. Sie sehen die Sortierung anhand eines kleinen Pfeils, der in der Spaltenüberschrift erscheint. Durch diese Darstellung können Sie auffällige Werte identifizieren.

- Sie interessieren sich für einige auffällige Werte? Markieren Sie die näher zu betrachtenden Formate einfach in der Tabelle. Bedenken Sie, dass sie für weitere Auswertungen die Auswahl wieder aufheben sollten.
- Zur Betrachtung der Daten, in denen die auffälligen Werte vorkommen, klicken Sie auf den Button „**Ausgangsdaten anzeigen**“ und Sie können so den konkreten Datenbestand kontrollieren.



Wertevorkommen			
Tabellenname	Spaltenname	Wert	Anzahl
			5.066
customer	marital_status	M	2.534
customer	marital_status	S	2.531
customer	marital_status		0
			1

Abbildung 27: Wertevorkommen 2

Anzahl Werte in %

Zum Aufruf der Auswertung „Anzahl Werte in %“ doppelklicken Sie in der Navigationsleiste auf „Anzahl Werte in %“. Die Auswertung zeigt in einem Balkendiagramm an wie viele Werte oder wie viele verschiedenen Werte sich in den einzelnen Spalten befinden. Um eine Berechnung der Auswertung zu erreichen, müssen Sie zuerst eine Tabelle selektieren. Um zwischen den Berechnungen „Anzahl Werte“

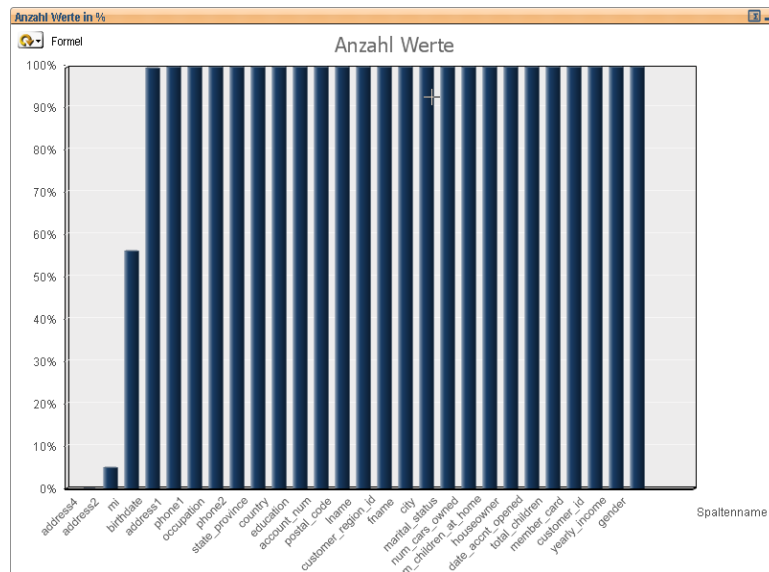


Abbildung 28: Anzahl der Werte

und „Anzahl eindeutiger Werte“ zu wechseln

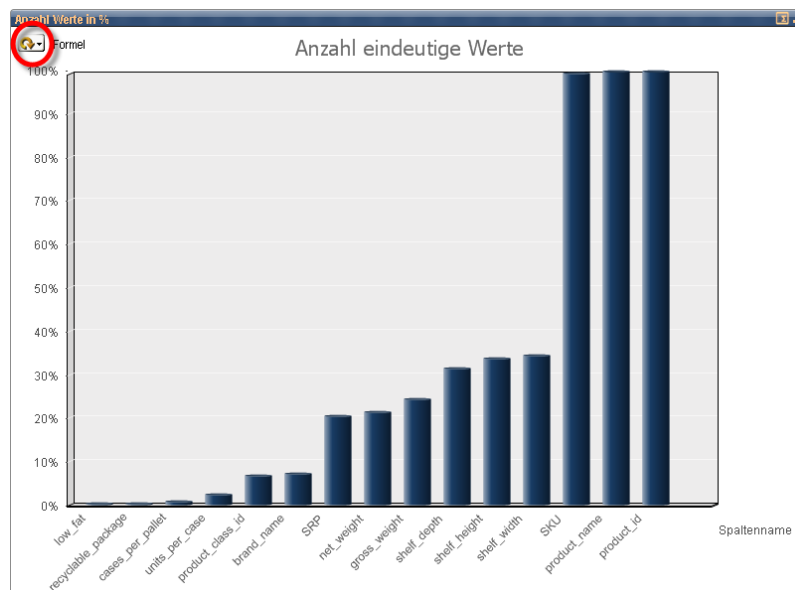


Abbildung 29: Anzahl der verschiedenen Werte

nutzen Sie bitte den dargestellten Button  **Formel**, welchen Sie innerhalb der

Auswertung rechts oben finden.

Mengenverteilung in %

Zum Aufruf der Auswertung „Mengenverteilung in % “ doppelklicken Sie in der Navigationsleiste auf „**Mengenverteilung in %**“.

Die Auswertung „Mengenverteilung in %“ zeigt in einem Kreisdiagramm an, wie die Datenmengen in den einzelnen Tabellen und Spalten verteilt sind.

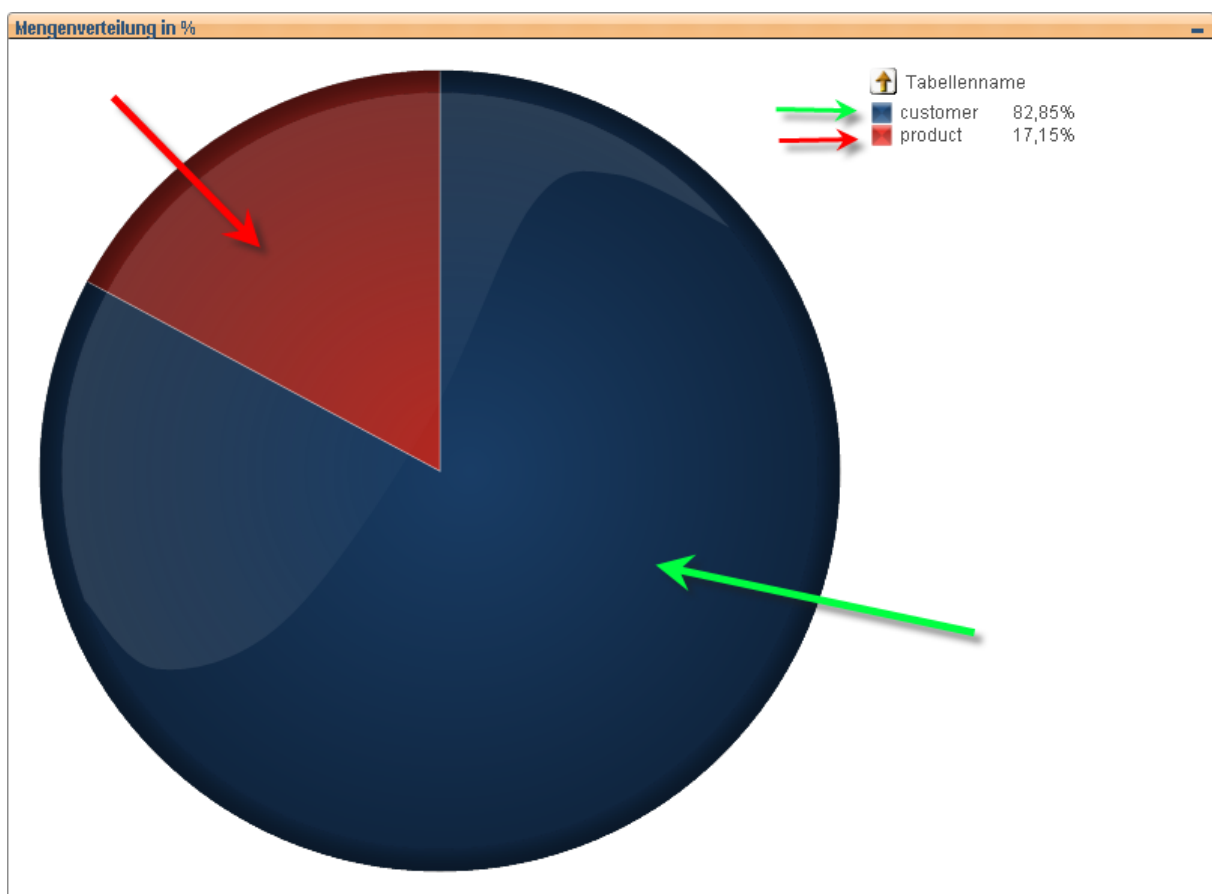


Abbildung 30: Mengenverteilung in %

Indem Sie einen durch die Pfeile angezeigten Bereiche klicken, gelangen Sie in die Ansicht der Mengenverteilung auf Spaltenebene, für die durch den Klick ausgewählte Tabelle.

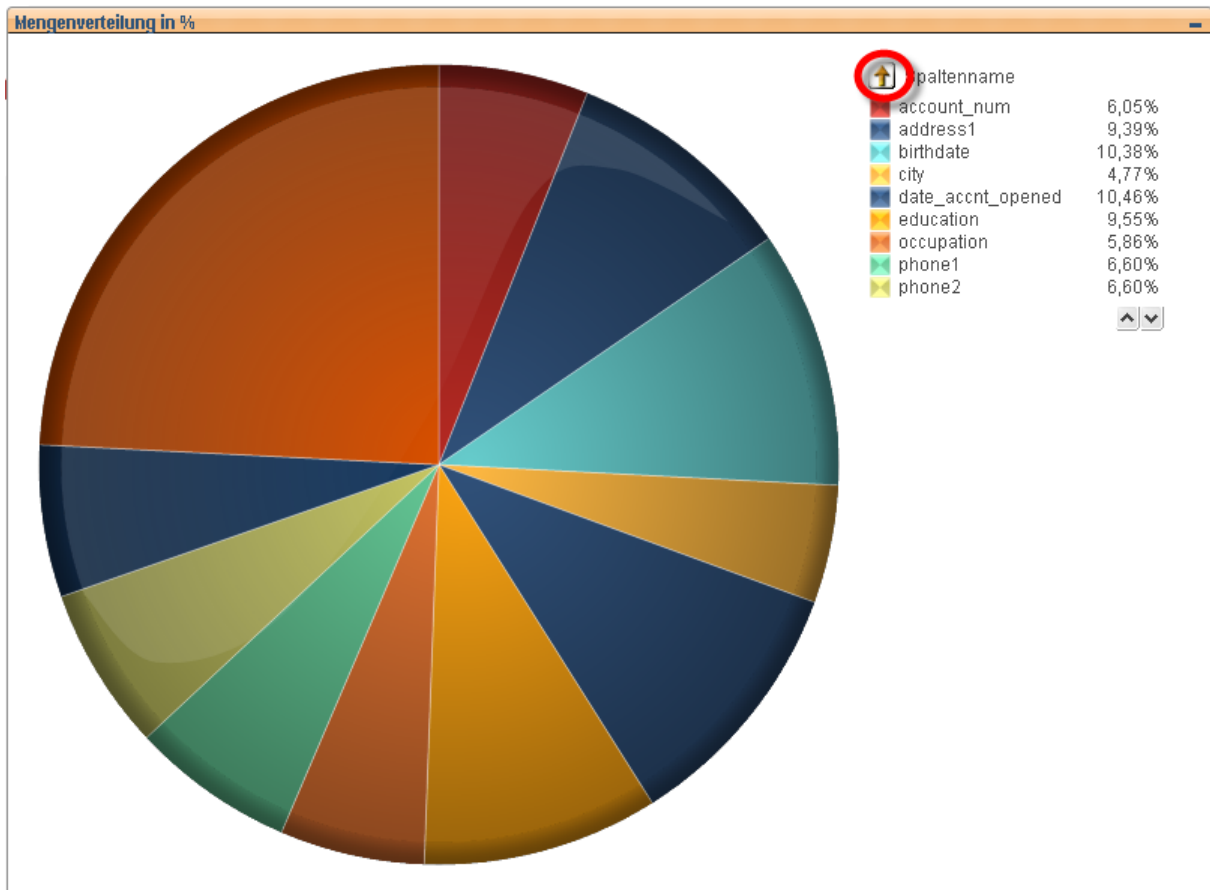



Abbildung 31: Mengenverteilung in %

Nutzen Sie bitte den dargestellten Button  Spaltenname , welchen Sie innerhalb der Auswertung rechts oben finden, um zu der Tabellenansicht zurück zu gelangen.

Null-Werte / Verschiedene Werte

Zum Aufruf der Auswertung „Auswertung Null / Verschiedene Werte“ doppelklicken Sie in der Navigationsleiste auf den entsprechend beschrifteten Button. Die Auswertung zeigt die Häufigkeit des Auftretens von Null-Werten je Spalte einer Tabelle, sowie die Anzahl von untereinander verschiedenen Werten, die in einer Spalte vorkommen.

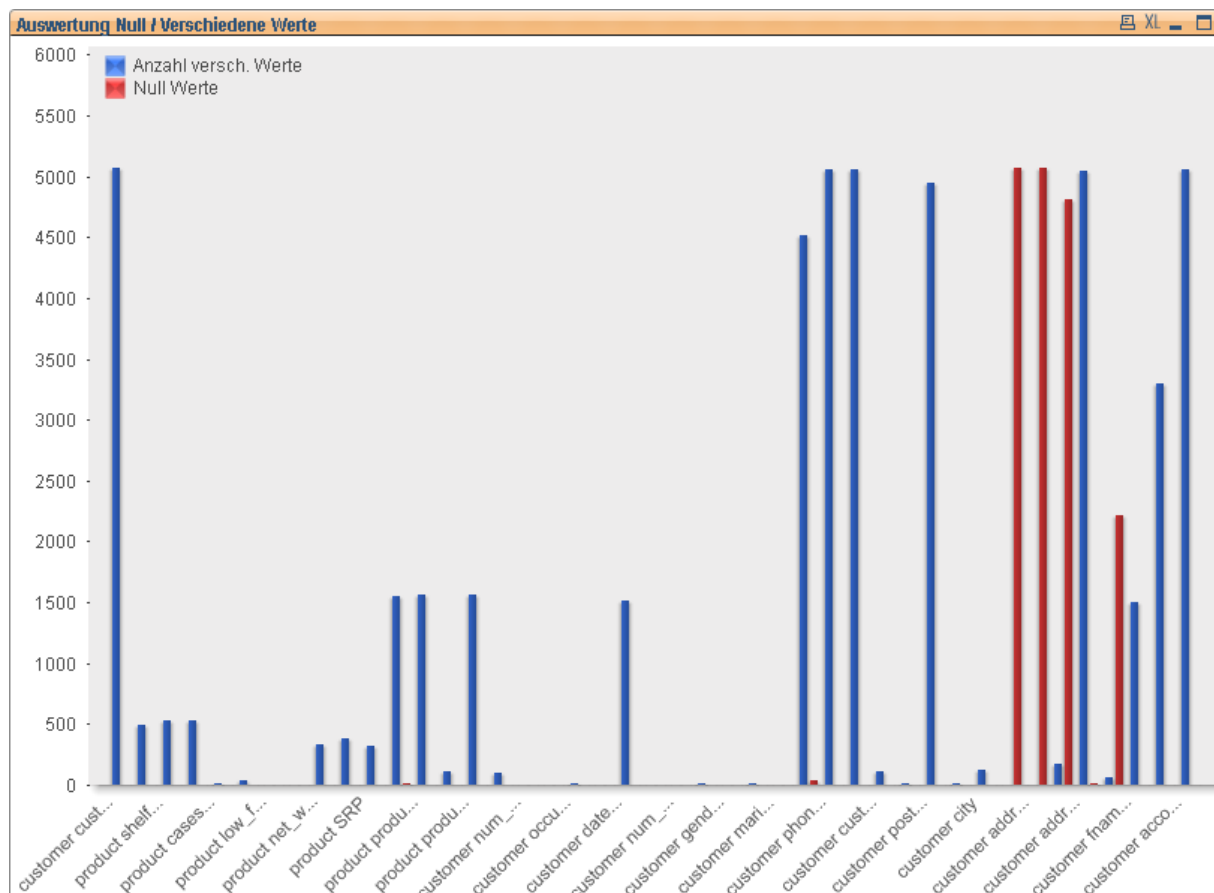


Abbildung 32: Auswertung Null/Verschiedene Werte

Ausgangsdaten anzeigen / Auswertungen anzeigen

Über den Button „**Ausgangsdaten anzeigen**“ gelangen Sie zu den vollständigen Datensätzen für die von Ihnen ausgewählten auffälligen Daten.



Abbildung 33: Ausgangsdaten anzeigen

Ausgangsdaten								
Tabellenname	PK	Spaltenname	account_num	address1	address2	address3	address4	birthdate
customer		1072	10610213300	7996 Ronda Ct.	-	-	-	1954-11-12
		1189	11401228900	3069 Yolanda	-	-	-	1923-10-19
		1197	11429612000	4524 Ferndale ...	-	-	-	1912-08-28
		1286	11993433833	174 Kim Court	-	-	-	1916-11-06
		1660	14502806663	1748 Bermad D...	-	-	-	1942-04-19
		1762	15097956657	6930 Lake Nadi...	-	-	-	1953-10-12
		2042	17182741200	9743 Post Road	#31	-	-	1965-07-20
		2171	18026043800	6384 Euclid Ave.	-	-	-	1973-07-07
		2617	21363538600	2669 Black Poin...	-	-	-	1957-10-23
		2981	23975118994	5871 Bryce Drive	-	-	-	1929-08-18
		3153	25214503400	6469 Castlewood	Unit B	-	-	1923-11-03
		3161	25252837538	2612 Berry Dr	-	-	-	1931-12-06
		3819	30341841000	8798 Valley Ma...	-	-	-	1955-06-17
		4326	34062145352	2212 Palmer Rd	-	-	-	1920-09-15
		4803	37710223197	4460 Carob Way	-	-	-	1961-04-22
	4958	38845226800	2916 Hillsborou...	-	-	-	1915-08-16	

Abbildung 34: Ausgangsdaten

Über den Button „Auswertungen anzeigen“ gelangen Sie wieder zurück zu den oben beschriebenen Analysen.

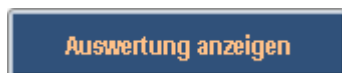


Abbildung 35: Auswertungen anzeigen

Ant. Werte nicht leer



Abbildung 36: Ant. Werte nicht leer

Diese Darstellung zeigt den prozentualen Anteil der Null-Werte des jeweils ausgewählten Datenbestandes und die nur mit Leerzeichen gefüllten Werte als sogenannte „leere“ Werte an.

Treffen Sie mit diesem Diagramm schnell eine Aussage, wie gut bestimmte Datenfelder (z.B. die Telefonnummer im Kundenbestand) gepflegt wurden.

Ant. Eindeutige Werte

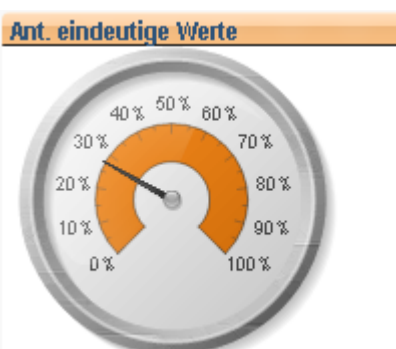
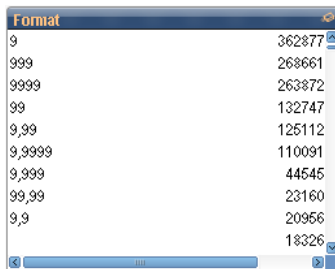


Abbildung 37: Ant. eindeutige Werte

Differieren die Werte in einer Spalte oder sind sehr viele identische Werte zu finden? Anhand dieser Grafik können Sie Aussagen zur Redundanz der Daten treffen und damit zusammenhängende Datenqualitätsprobleme aufspüren, die sich in den operativen Bereichen Ihres Unternehmens wie dem

Kundenmanagement niederschlagen könnten.

Listbox „Format“



Format	Häufigkeit
9	362877
999	268661
9999	263872
99	132747
9,99	125112
9,9999	110091
9,999	44545
99,99	23160
9,9	20956
	18326

Die Listbox zeigt die vorkommenden Muster der getroffenen Auswahl und deren Häufigkeiten absteigend geordnet an. Zahlen werden durch „9“ symbolisiert, Buchstaben durch „a“, alle anderen Zeichen sind unverändert. „Formatausreißer“ können so gezielt erkannt werden.

Abbildung 38: Format

Listbox „Wert“



Wert	Häufigkeit
0	125982
3	85369
4	55300
2	40997
1	14223
12	12501
8	12310
19	12059
17	12025
13	11906

Alle in der jeweils selektierten Auswahl vorkommenden Werte sowie die Häufigkeit ihres Auftretens werden in der Listbox angezeigt. Eine Prüfung der Werte ist sehr einfach möglich, wenn z.B. nur eine definierte Menge von Werten auftreten darf. Auch redundante Werte, z.B. Telefonnummern können leicht erkannt werden.

Abbildung 39: Wert

3. Übungsbeispiele

3.1. Analyse Formatvorkommen von Postleitzahlen für Kunden aus den USA

1. Zunächst müssen alle Datensätze, die das Land USA enthalten aus den Kundendaten identifiziert werden. Folgen Sie hierfür der Anleitung in Schritt 2-5. Die eigentliche Analyse beginnt ab Schritt 6.
2. Wählen Sie die Tabelle mit den Kundenstammdaten „customer“ aus der Listbox „Tabellen“ aus.
3. Klicken Sie auf die Titelleiste der Listbox Spalten und tippen sie auf der Tastatur das Wort *country*. Wählen sie die Spalte „country“.

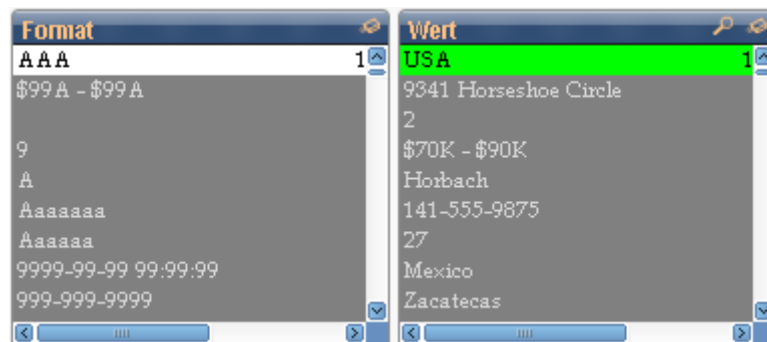


Abbildung 40: Selektion USA

4. Selektieren sie in der Listbox „Werte“ den Wert *USA*. Durch diese Einschränkung ist nun keine direkte Formatanalyse der Postleitzahlen möglich, wie Abbildung 40 verdeutlicht.
5. Um nun alle Datensätze der Kunden aus den USA für die Postleitzahlenauswahl zu fixieren und auf das Format der Postleitzahlen zugreifen zu können, nutzen wir die Primärschlüssel als eindeutiges Attribut. Klicken Sie in der Multibox mit der rechten Maustaste auf Primärschlüssel und wählen sie im Kontextmenü **„Wählbare Werte auswählen“**. Wiederholen Sie den Vorgang und klicken Sie im Kontextmenü auf **„Auswahl sperren“**.
6. Nun kann die Auswahl der Tabelle und des Landes wieder aufgehoben werden. Klicken Sie dazu **„Auswahl aufheben“** in der Symbolleiste. Die Datensätze mit dem Land USA bleiben durch die Sperrung der Primärschlüssel erhalten.
7. Wählen Sie jetzt in der Listbox „Spalten“ die Spalte postal_code, wie in

der
Abbildung 41.

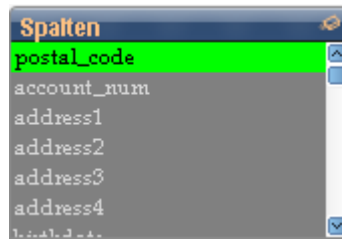


Abbildung 41: Auswahl Postleitzahlen

8. Klicken Sie auf „**Formatvorkommen**“ und beginnen Sie Ihre Analyse.

Formatvorkommen				
Tabellenname	Spaltenname	Format	Anzahl	%
			3.626	100,00%
customer	postal_code	99999-9999	1	0,03%
customer	postal_code	99999 Aaa Aaaaaaaaa	1	0,03%
customer	postal_code	A A A	1	0,03%
customer	postal_code	aaaaaa	1	0,03%
customer	postal_code	99999999	1	0,03%
customer	postal_code	9999999999	1	0,03%
customer	postal_code	9999999	2	0,06%
customer	postal_code	9999	10	0,28%
customer	postal_code	99999	3.608	99,50%

Abbildung 42: Analyse Formatvorkommen für Postleitzahlen USA

9. Sofern Sie weitere Analysen vornehmen möchten, vergessen Sie nicht, die Sperrung über den Button „**Freigeben**“ aufzuheben und die Auswahl über den Button „**Auswahl aufheben**“ zu löschen.

3.2. Identifizierung von möglichen Schlüsselattributen

1. Wählen Sie die Tabelle mit den Kundenstammdaten „customer“ aus der Listbox „Tabellen“ aus.
2. Sofern die Analyse „Auswertungen allgemein“ nicht geöffnet ist, klicken Sie den Button „**Auswertungen Anzeigen**“ und betrachten Sie die Spalte „Anz. Versch. Werte“.
3. Unter der Listbox „Tabellen“ finden Sie die Anzahl der Datensätze in der Tabelle „customer“. Handelt es sich um ein Schlüsselattribut beziehungsweise um einen Schlüsselkandidaten, müssen die Werte der gesuchten Spalte eindeutig sein. Das bedeutet jeder Wert darf nur genau einmal vorkommen.
4. In unserem Beispiel kommt die Spalte „customer_id“ als Schlüsselattribut in Frage, da hier die Anzahl der verschiedenen Werte der Anzahl der Datensätze in der Tabelle „customer“ entspricht.

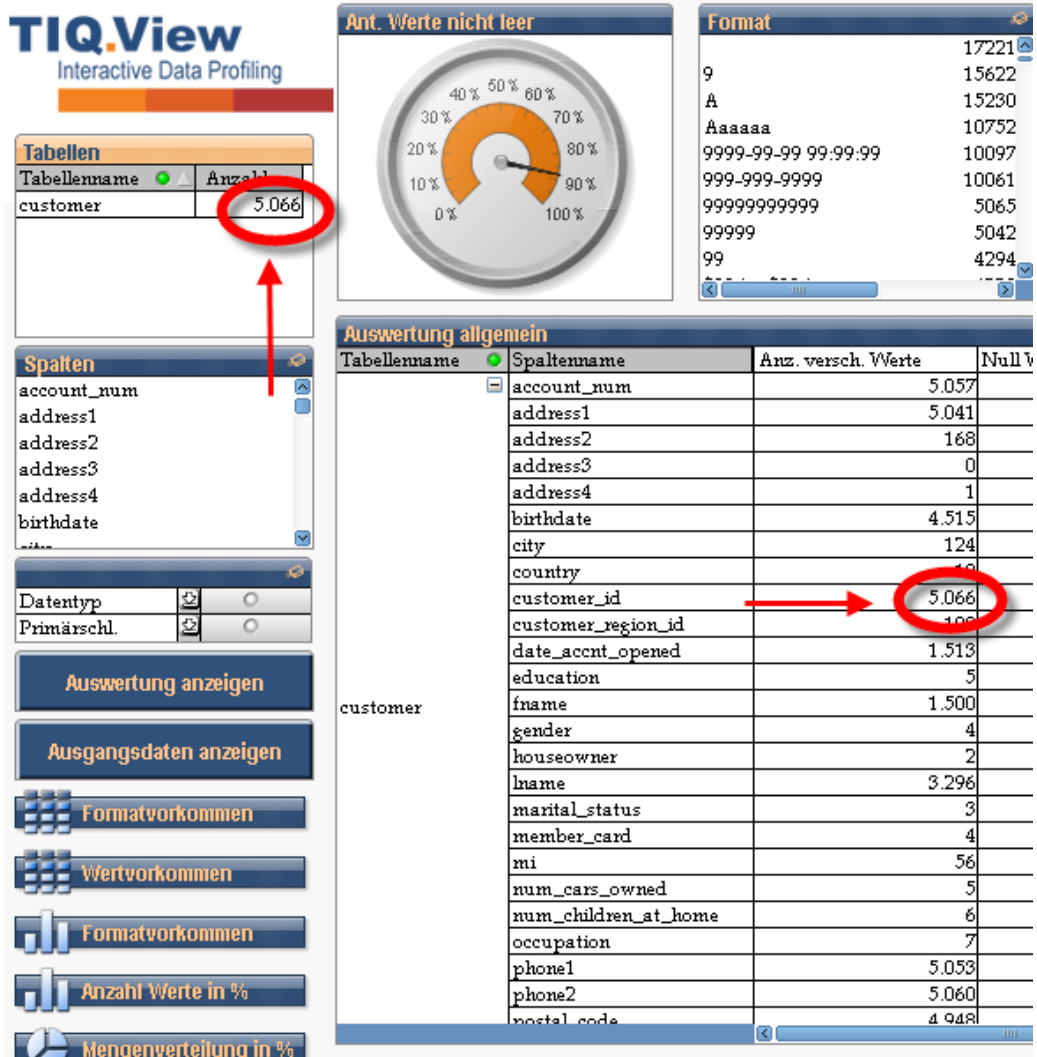


Abbildung 43: Identifizierung von Schlüsselattributen

3.3. Wertevorkommen

1. Wählen Sie in der Listbox „Spalten“ die Spalte „birthdate“ aus
2. Danach sollten Sie den Button „**Wertevorkommen**“ öffnen und die Werte nach der Größe sortieren: Dort können Sie Ausreißer-Werte z.B. aus den 16. Jahrhundert leicht identifizieren.
3. Selektieren Sie die ungewöhnlichen Wertevorkommen und klicken Sie auf „**Ausgangsdaten anzeigen**“, um die entsprechenden Datensätze anzuzeigen.



Abbildung 44: Beispiel 3 Wertevorkommen

- Die identifizierten falschen Datensätze können z.B. nach Excel exportiert oder ausgedruckt werden. Für den Excel-Export finden Sie ein Symbol „XL“, für Drucken ein kleines Drucker-Symbol in der Titelleiste der Analyse „Auswertungen“.

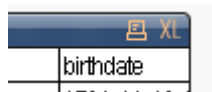
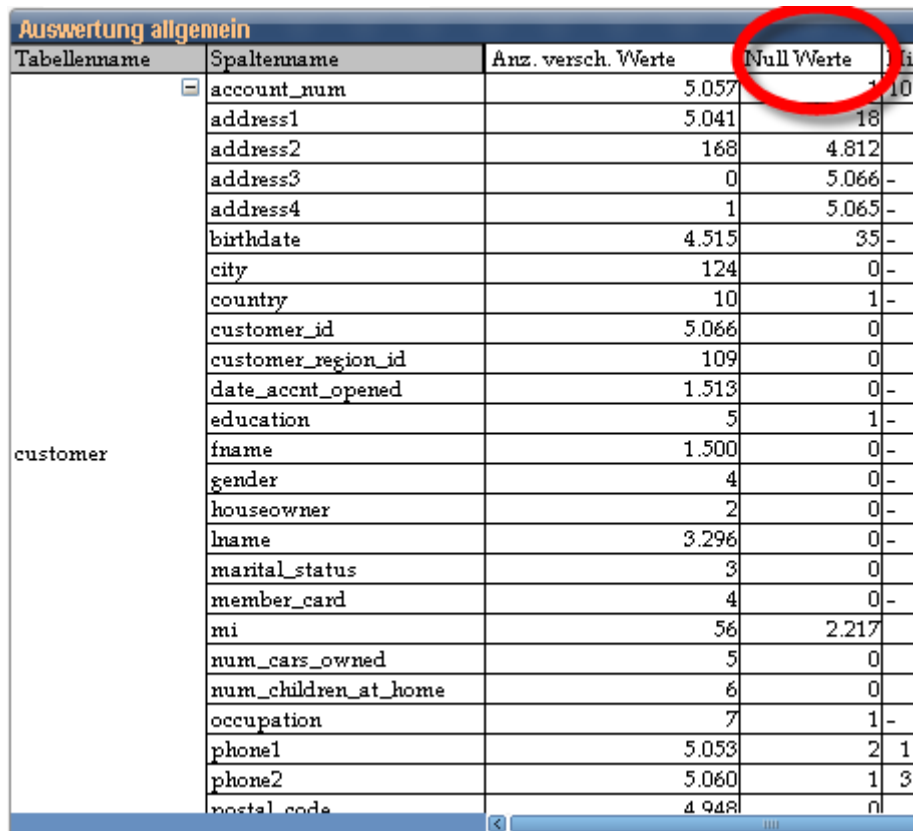


Abbildung 45: Excel-Export und Drucken

3.4. Nullwert-Analyse

1. In der Analyse „Auswertungen allgemein“ kann in der Spalte „Null Werte“, die Verteilung der Null-Werte analysiert werden.
2. Selektieren Sie auffällige Nullwert-Vorkommen und klicken Sie auf „Ausgangsdaten anzeigen“, um die entsprechenden Datensätze anzuzeigen.



Tabellenname	Spaltenname	Anz. versch. Werte	Null Werte	...
customer	account_num	5.057	10	
	address1	5.041	18	
	address2	168	4.812	
	address3	0	5.066	
	address4	1	5.065	
	birthdate	4.515	35	
	city	124	0	
	country	10	1	
	customer_id	5.066	0	
	customer_region_id	109	0	
	date_accnt_opened	1.513	0	
	education	5	1	
	fname	1.500	0	
	gender	4	0	
	houseowner	2	0	
	lname	3.296	0	
	marital_status	3	0	
	member_card	4	0	
	mi	56	2.217	
	num_cars_owned	5	0	
num_children_at_home	6	0		
occupation	7	1		
phone1	5.053	2	1	
phone2	5.060	1	3	
postal_code	4.948	0		

Abbildung 46: Nullwertanalyse am Beispiel

TIQ Solutions GmbH
Industriestraße 85-95
D-04299 Leipzig

tel: +49 (0) 341.35 59 03 00
fax: +49 (0) 341.35 59 03 25

email: info@tiq-solutions.de
web: www.tiq-solutions.de