Explorative Datenanalyse mit SAS Visual Analytics unter SAS Viya

Christoph Frank
HMS Analytical Software GmbH
Rohrbacher Straße 26
69115 Heidelberg
Christoph.frank@analytical-sotware.de

Zusammenfassung

Dieser Beitrag demonstriert die Report-Erstellung in SAS Visual Analytics unter SAS Viya zur explorative Datenanalyse. Der entsprechende Workflow vom Einlesen der Daten über die Datenaufbereitung und das Report-Design bis zur Auswertung der Daten wird anhand eines Beispiels mit fiktiven Bankdaten vorgeführt.
Die gewünschten Zielgrößen lassen sich schnell und einfach in interaktiven Reports visualisieren, die durch individuelle Eingaben des Report-Empfängers in Echtzeit angepasst werden können, um die Daten auf bestimmte Muster zu untersuchen und neue Erkenntnisse zu gewinnen.

Schlüsselwörter: SAS Visual Analytics, SAS Viya, SAS Cloud Analytic Services (CAS), SAS Studio, SAS Data Studio, Data Views

1 Über SAS Visual Analytics

1.1 Was ist SAS Visual Analytics?

SAS Visual Analytics bietet die Möglichkeit, Informationen aus Daten nutzerfreundlich zu visualisieren und zu analysieren. Durch die Erstellung interaktiver Berichte können Muster und Zusammenhänge in großen Datenmengen durch eine leistungsstarke In-Memory-Verarbeitung in Echtzeit dargestellt werden.
SAS Visual Analytics ist eine Web-Anwendung (html5), das heißt, dass kein lokaler Client installiert werden muss. Die Berichterstellung bzw. Visualisierung ist Point-and-Click orientiert. Schritte zur Datenaufbereitung und -analyse können aber auch im SAS Studio programmiert werden. Insbesondere sind mit dem SAS Scripting Wrapper for Analytics Transfer (SWAT) Package Schnittstellen zu anderen Programmiersprachen wie Python oder R verfügbar.

1.2 nstallation von SAS Visual Analytics

In SAS Visual Analytics können unterschiedliche Benutzerrollen vom SAS-Administrator über den Report-Designer und Analysten bis hin zum reinen Report-

Viewer (Empfänger) ohne technisches Hintergrundwissen eingerichtet werden. Die SAS Viya Installation umfasst je nach Lizenzierung verschiedene Komponenten.
Im Rahmen dieser Demonstration nutzen wir die Anwendungen SAS Studio, SAS Data Studio und SAS Visual Analytics. Dabei wird die SAS Viya Plattform auf einer virtuellen Maschine mit 128 GB Arbeitsspeicher in einer Microsoft Azure Cloud betrieben.

2 Beispieldaten und zu untersuchende Fragestellung

Die vorliegenden Beispieldaten umfassen ca. 4 Millionen Zeilen und 34 Spalten mit fiktiven Bankdaten, insbesondere Gewinne und Verluste für verschiedene Bank- und Versicherungs-Produkte, die über eine Zeitspanne von 7 Jahren an verschiedenen Standorten verkauft wurden. Das Produktportfolio kann in verschiedenen Ebenen in Kategorien und Unterkategorien zusammengefasst werden. Beispielsweise enthält die Geschäftssparte *Insurance* die Produktlinien *Vehicle*, *Life* und *Property*. Außerdem lässt sich der Verkauf der Produkte bzw. die Transaktionen über Regionen und Zeiträume aggregieren. Das Bundesland Baden-Württemberg ist beispielsweise untergliedert in die Gebietsfilialen Stuttgart und Freiburg, unter denen wiederum verschiedene Niederlassungen zusammengefasst werden.

Im Rahmen dieses Beitrags soll demonstriert werden, wie SAS Visual Analytics eingesetzt werden kann, um den relativen Gewinn von Produkten und Produktkategorien auf zeitliche und regionale Unterschiede zu untersuchen.

3 Vorgehen bei der explorativen Datenanalyse in SAS Visual Analytics

3.1 Arbeiten mit SAS Visual Analytics

Der SAS Drive (ehemals SAS Home) ist die Benutzeroberfläche, auf der alle SAS Viya Anwendungen und gespeicherten Inhalten übersichtlich dargestellt werden und einfach zugänglich sind (siehe Abbildung 1).

Im SAS Drive findet man unter *Daten verwalten* eine einfach zu bedienende Benutzeroberfläche, um Daten verschiedener Formate für die Auswertung in SAS Visual Analytics zu importieren. Die importierten Daten werden im SASHDAT-Format in-Memory verarbeitet und können auch in diesem Format gespeichert werden, um sie für zukünftige Auswertungen schneller in den Hauptspeicher laden zu können.

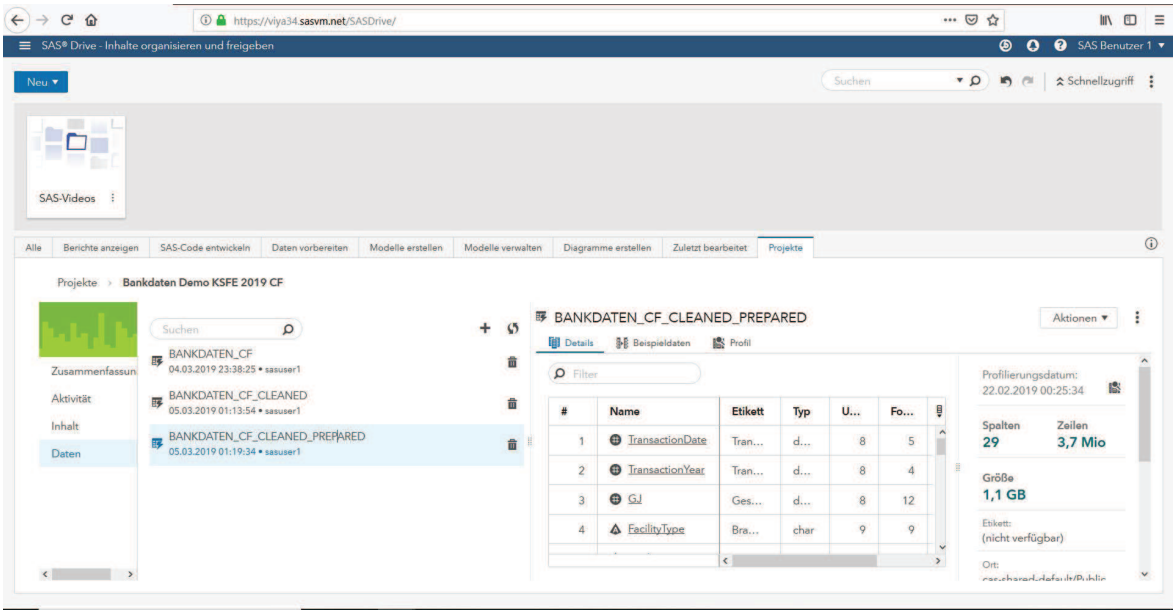

Abbildung 1: SAS Drive – Übersicht aller Anwendungen und Inhalte

Um die importierten Daten für die Reporterstellung aufzubereiten, kann das SAS Data Studio genutzt werden, welches auch für Anwender ohne Programmierkenntnisse umfangreiche Möglichkeiten zur Datenaufbereitung bietet. Darunter sind einfache Transformationen wie das Zerlegen von String-Variablen, die Berechnung neuer Variablen oder das Konvertieren von Variablentypen. Es können aber auch Filter erstellt werden, um nur Zeilen zu berücksichtigen, die bestimmte Bedingungen erfüllen. Darüber hinaus können Datensätze transponiert werden oder mehrere Datenquellen miteinander verbunden werden.

Falls eine spezifische Datenaufbereitung benötigt wird, die mit Hilfe des SAS Data Studios nicht umsetzbar ist, kann ein entsprechendes SAS-Skript im SAS Studio erstellt werden. Unter SAS Viya haben wir die Möglichkeit, sowohl mit klassischen SAS BASE Sprachelementen zu arbeiten als auch CAS (Cloud Analytic Server) Syntax zu nutzen bzw. Operationen auf SASHDAT-Dateien auszuführen, um den Performancegewinn durch parallele Verarbeitung zu nutzen.

Wenn die Datenaufbereitung abgeschlossen ist, können die Berichte zur explorativen Analyse im Report-Designer erstellt werden (siehe Abbildung 2). Hier findet man eine Vielzahl von Visualisierungsobjekten wie Tabellen und Grafiken.

Die Spalten der Quelldatei werden automatisch in kategorielle und numerische Variablen unterteilt. Außerdem können Variablen als geografische Variablen klassifiziert werden, die entweder aufgrund ihres Namens bzw. Codes oder durch explizite Zuordnung von Höhen- und Breitengrad-Variablen von entsprechenden Visualisierungsobjekten (z.B. Landkarten) dargestellt werden können.

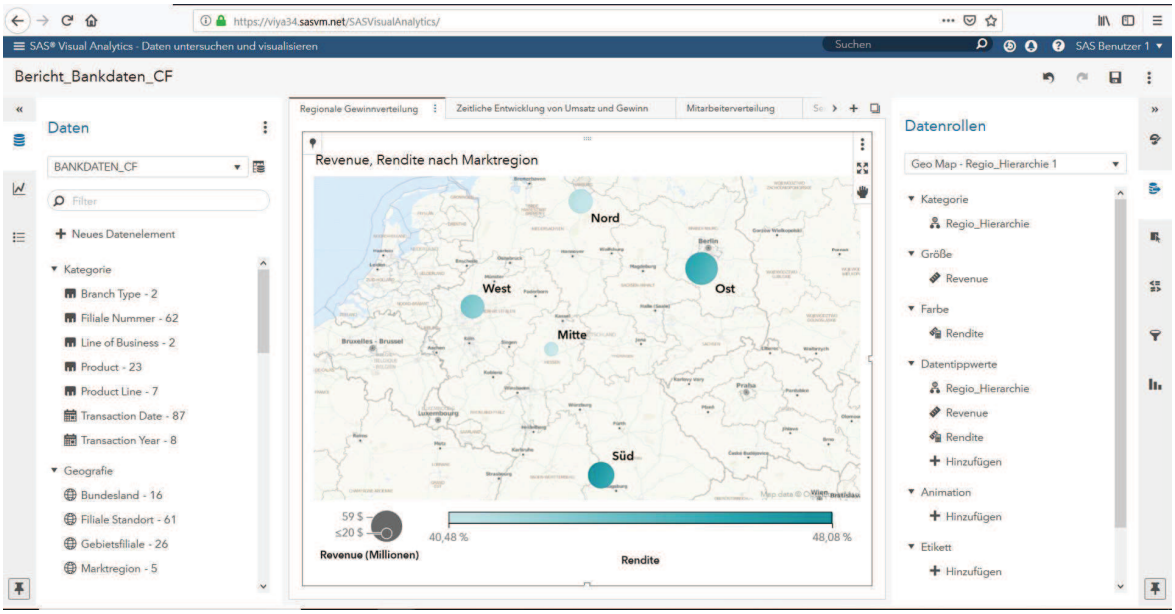

Abbildung 2: Der Report-Designer von SAS Visual Analytics

Enthalten die zugrundeliegenden Daten Informationen in Form von Kategorien und Unterkategorien, das heißt, wenn Spalten enthalten sind, deren Werte jeweils Werte anderer Spalten zusammenfassen, ist das Erstellen von Hierarchien eine nützliche Möglichkeit, diesen Zusammenhang in interaktiven Reportobjekten zu veranschaulichen.

Wenn die Anwendungen SAS Visual Statistics und SAS Data Mining and Machine Learning ebenfalls lizenziert wurden, findet man im Report-Designer auch analytische Werkzeuge, um klassische statistische Modelle zu erstellen oder moderne Machine Learning Algorithmen anzuwenden, um Zusammenhänge innerhalb der Daten zu verifizieren.

3.2 Workflow dieser Demonstration

Wir nutzen den im vorherigen Abschnitt beschriebenen Import-Wizard, um unseren Beispieldatensatz, der im klassischen SAS Dataset-Format (SAS7BDAT) vorliegt, zu importieren und erstellen daraus eine SASHDAT-Datei. Diese wird physisch gespeichert, muss aber stets in den Hauptspeicher geladen werden, um damit arbeiten zu können.

Eine erste Visualisierung des unverarbeiteten importierten Datensatzes offenbarte fehlerhafte Daten. Die Zuordnung von Längen- und Breitengraden zu den Filialstandorten war nicht eindeutig, wie Abbildung 3 verdeutlicht.

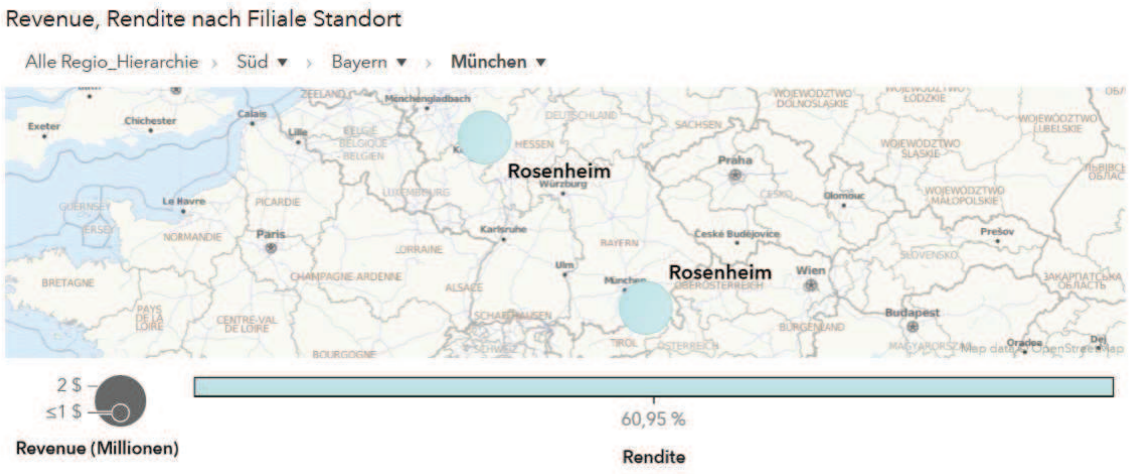

Abbildung 3: Eine erste Visualisierung der Ortsdaten zeigte fehlerhafte Werte

Die Datenfehler sollen nun durch das im Folgenden dargestellte Programm im SAS Studio bereinigt werden.

```
/*************** Unsaubere Daten entfernen **********************/

***** CAS session starten;
cas demo_session;

***** Widersprüchliche Ortsdaten entfernen;
proc fedsql sessref=demo_session;
  create table public.bankdaten_CF_cleaned {options replace=true} as
  select a.* from public.bankdaten_CF a
  left join (select FilialeOrt
             from public.bankdaten_CF
             group by FilialeOrt
             having count(distinct FilialeOrt_lat)>1 or
                    count(distinct FilialeOrt_long)>1) b
     on a.FilialeOrt = b.FilialeOrt
     where b.FilialeOrt is null;
quit;


***** Fehlerhafte Zuordnung korrigieren;
libname mycaslib cas caslib=public;
data mycaslib.bankdaten_CF_cleaned;
  set mycaslib.bankdaten_CF_cleaned;
  if FilialeOrt = 'Bautzen' and Gebietsfiliale = 'Magdeburg' then
  do;
    FilialeOrt = Gebietsfiliale;
    FilialeOrt_long = Gebietsfiliale_long;
    FilialeOrt_lat = Gebietsfiliale_lat;
  end;
run;

****** bereinigte Daten speichern;
```

```
proc casutil incaslib="public" outcaslib="public";
  promote casdata="bankdaten_CF_cleaned";
quit;
proc casutil;
  save casdata="bankdaten_CF_cleaned"
       incaslib="public"
       outcaslib="public"
       casout="bankdaten_CF_cleaned" replace;
run;
```

Sowohl die PROC FEDSQL Prozedur als auch der DATA STEP werden von der CAS Engine parallel ausgeführt, was zu einem erheblichen Performancegewinn verglichen mit klassischen SAS Prozeduren führt.

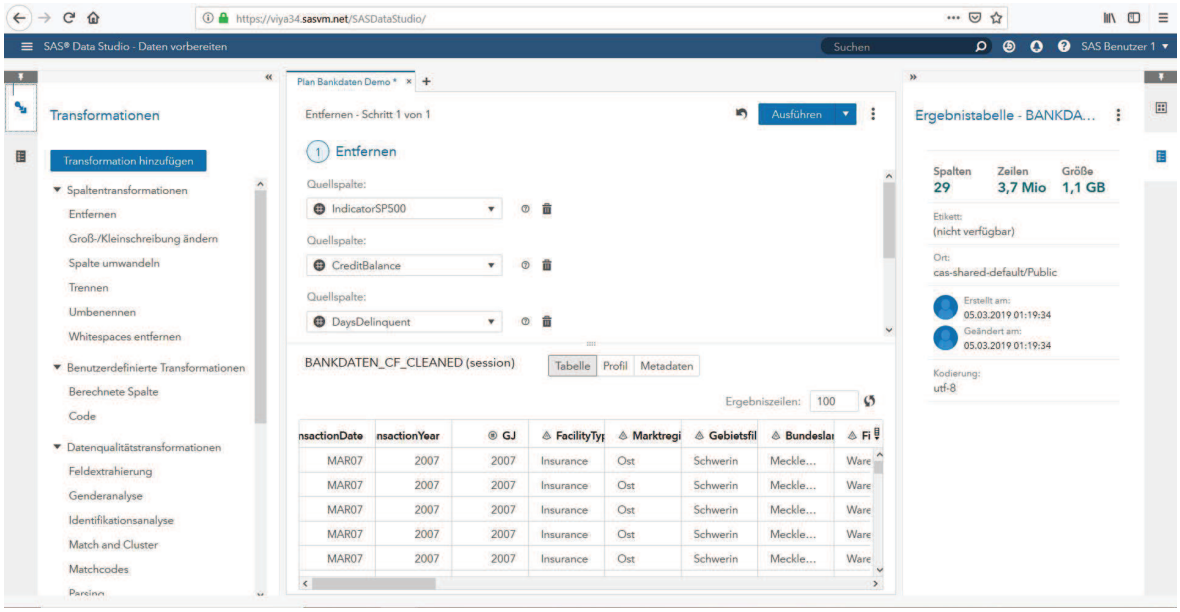

Abbildung 4: SAS Data Studio zur Point-and-Click Datenaufbereitung

Alle notwendigen Datenaufbereitungsschritte in dieser Demonstration hätten auch in obigen SAS-Programm umgesetzt werden können, aber wir nutzen zu Vorführzwecken das SAS Data Studio, um überflüssige Variablen zu entfernen (siehe Abbildung 4).

Abschließend erstellen wir im Report-Designer (siehe Abbildung 5) mit Hilfe von passenden Grafiken wie einer Geo-Map und eines Kacheldiagramms einen Report, der dem Empfänger bei der Beantwortung der eingangs formulierten Fragestellung unterstützen soll.

In SAS Visual Analytics werden in Grafiken und Diagrammen die darzustellenden Maße aus den metrischen Variablen berechnet, die dabei standardmäßig über alle Gruppierungsebenen der zugrundeliegenden kategorischen Variablen hinweg auf dieselbe Art aggregiert werden. Das bedeutet, dass alle Zeilen des Eingabedatensatzes, die eine Gruppe in Bezug auf eine oder mehrere kategorischen Variablen bilden, zusammenge-

fasst werden und dass dann z.B. die Summe oder der Durchschnitt der metrischen Variable zur Visualisierung berechnet wird.

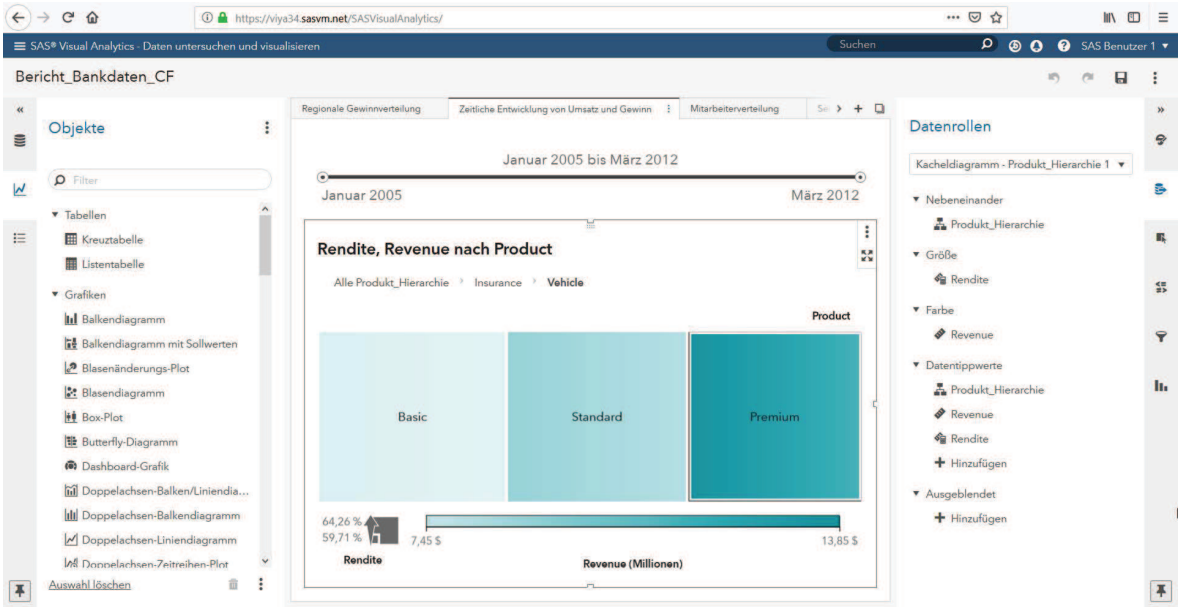

Abbildung 5: Erstellen von Reports im SAS Visual Analytics Report-Designer

Innerhalb des Report-Designers können neue Variablen erstellt werden. Insbesondere haben wir die Möglichkeit, aggregierte Maße zu definieren, die sich der jeweiligen Betrachtungsebene der Visualisierungsobjekte bzw. der Gruppierungsebene der im Objekt verwendeten Kategorien anpassen. Wir können somit das aggregierte Maß *Rendite* als prozentualen Anteil des Gewinns vom Umsatz definieren (siehe Abbildung 6), welches jeweils für die verschiedenen Betrachtungsebenen der Diagramme berechnet wird.

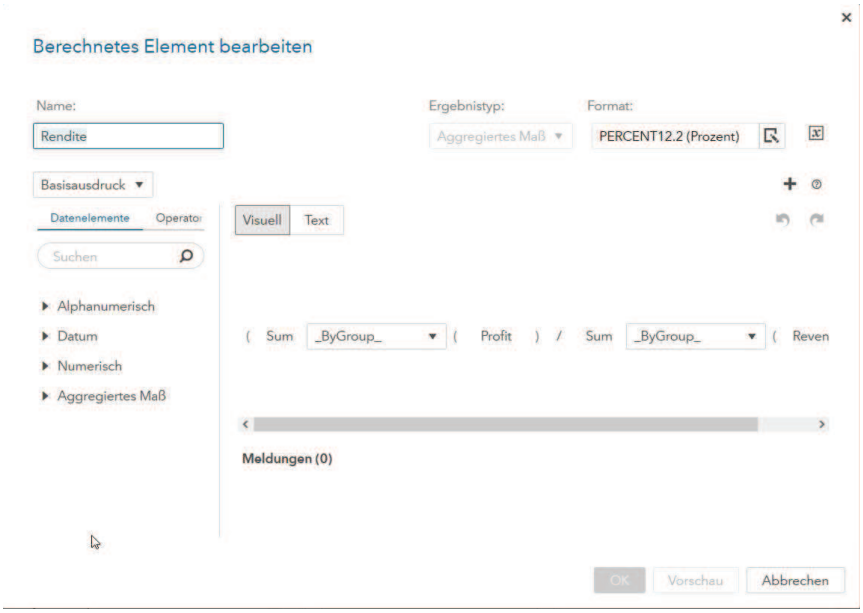

Abbildung 6: Definition des prozentualen Gewinns als aggregiertes Maß

Außerdem nutzen wir die Möglichkeit zur Definition von Hierarchien in SAS Visual Analytics und fassen kategorische Variablen über die Regionen bzw. die Produktsparten gemäß Tabelle 1 zusammen.

Tabelle 1: Zusammenfassen von Informationen über Regionen und Produktsparten in Hierarchien

	Regionale Hierarchie	Produkt Hierarchie
Ebene 1	Marktregion	Line of Business
Ebene 2	Bundesland	Product Line
Ebene 3	Gebietsfiliale	Product
Ebene 4	Filiale	

Unabhängig von der im Rahmen dieser Demonstration zu untersuchenden Fragestellung, eignet sich der Beispieldatensatz darüber hinaus zur Demonstration des *AggregateTable Operators*.
Mit dessen Hilfe kann ein neues Maß definiert werden und statt der oben erwähnten uniformen Aggregation kann explizit festgelegt werden, auf welcher Gruppierungsebene das neue Maß wie aggregiert werden soll. Abbildung 7 zeigt die entsprechende Benutzeroberfläche.

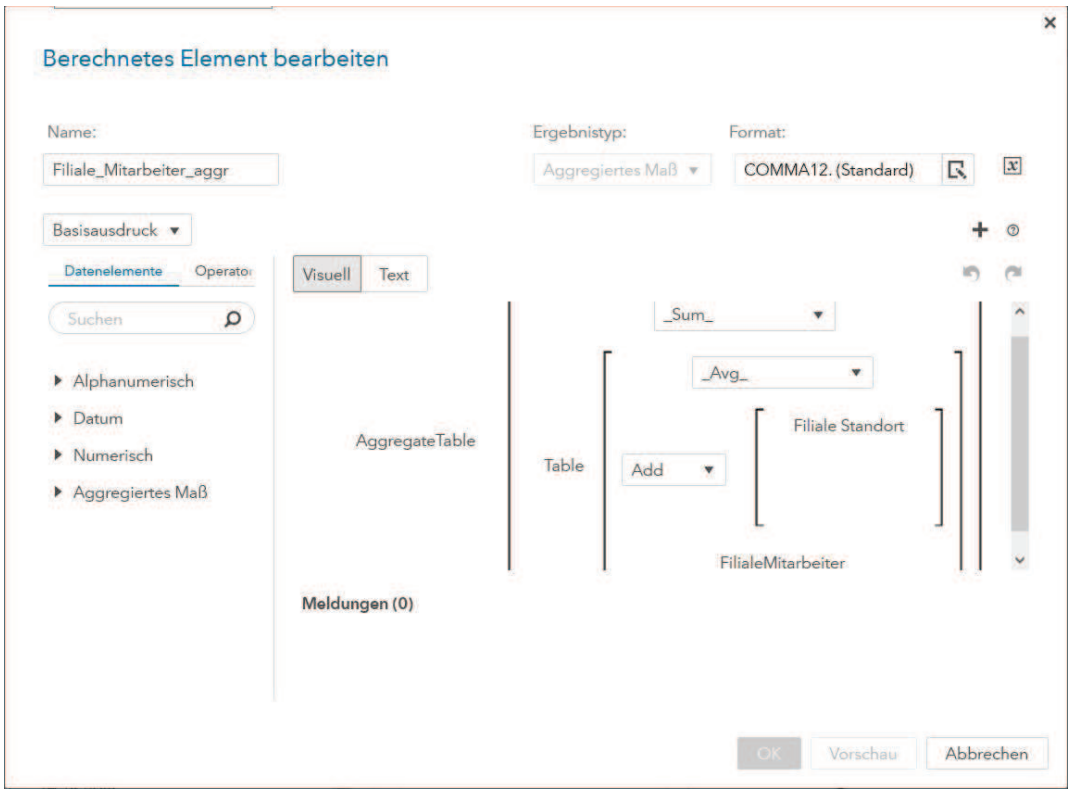

Abbildung 7: Definition einer neuen Variablen mittels des AggregateTable Operators

Hier wird das neue Maß *Filiale_Mitarbeiter_aggr* definiert. In unserem Beispieldatensatz entspricht jede Zeile einer Transaktion, die jeweils die kategorische Variable *Filiale Standort* und die metrische Variable *FilialeMitarbeiter* enthält. Soll nun in einem Visu-

alisierungsobjekt nach Filialen gruppiert werden, dann ergibt sich die Anzahl der Mitarbeiter aus dem eindeutigen Wert von *FilialeMitarbeiter* für *Filiale Standort*. Die Anzahl der Mitarbeiter einer Filiale kann auf der Gruppierungsebene *Filiale Standort* für das neue Maß *Filiale_Mitarbeiter_aggr* also als Durchschnitt (_Avg_ in Abbildung 7) von *FilialeMitarbeiter* modelliert werden. Für alle anderen Oberkategorien wie *Gebietsfiliale*, *Bundesland* etc. soll der Wert *FilialeMitarbeiter* zur Berechnung von *Filiale_Mitarbeiter_aggr* jedoch aufsummiert werden. In dem in Abbildung 8 dargestellten Report, wurde das neue Maß *Filiale_Mitarbeiter_aggr* und die Regionen-Hierarchie genutzt, um die Verteilung der Anzahl der Mitarbeiter in einem interaktiven Kreisdiagram abzubilden.

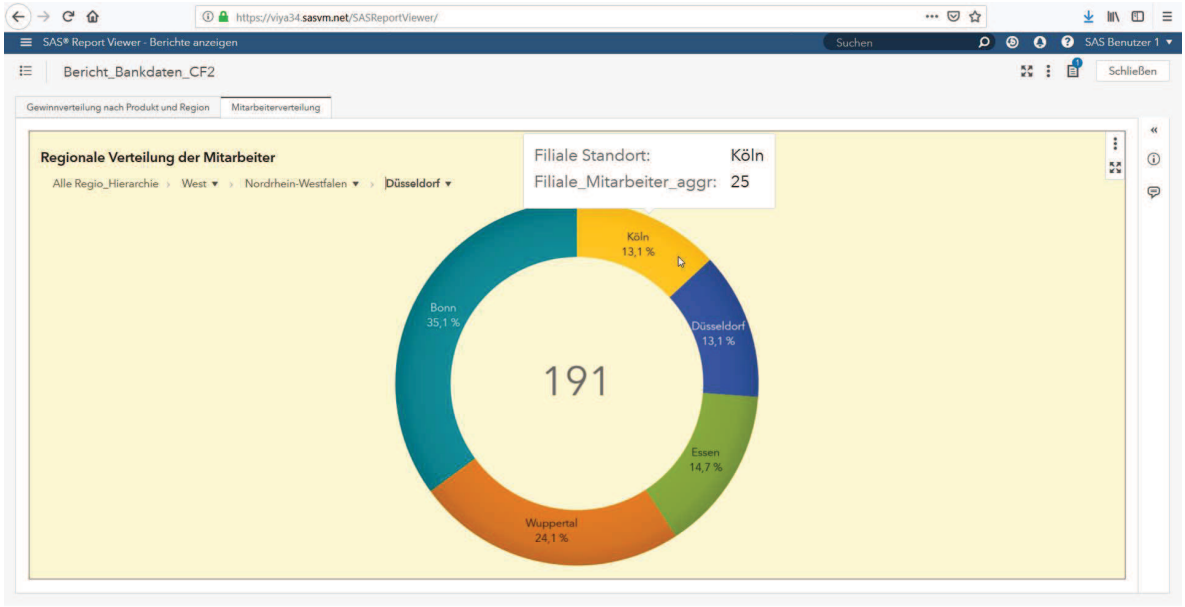

Abbildung 8: Kreisdiagramm über die regionale Verteilung der Mitarbeiterzahl

Um Hierarchien und aggregierte Maße nicht für jeden Report neu definieren zu müssen und auf strukturell gleiche Datensätze übertragen zu können, kann die Information in *Dataviews* gespeichert werden, die je nach Berechtigung einem Nutzer persönlich oder einer Gruppe von Nutzern zur Verfügung stehen.

3.3 Fertiger Report und Ergebnisse

Der fertige Report in Abbildung 9 enthält eine Geo-Map, die die geografischen Kategorien auf einer Landkarte abbildet und durch verschieden große und verschieden farbige Blasen die Maße *Revenue* und *Rendite* veranschaulicht. Diese Grafik wurde mit einem Kacheldiagram verknüpft, um die unterschiedlichen Ausprägungen von *Revenue* und *Rendite* in Abhängigkeit der Produktkategorien und der Standorte zu verdeutlichen.

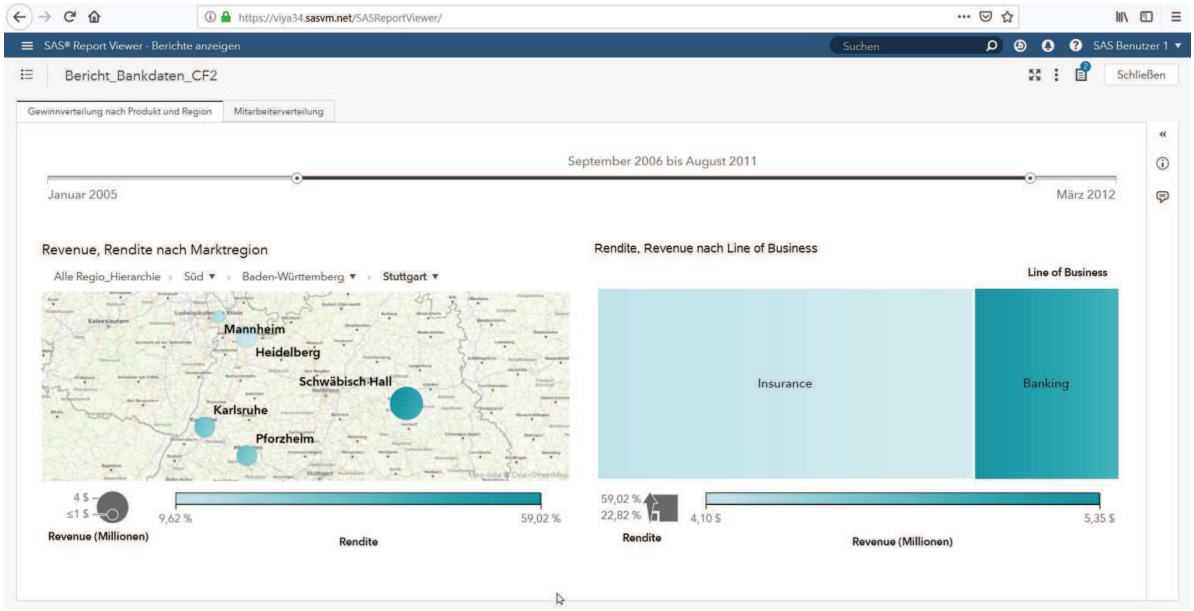

Abbildung 9: Fertiger interaktiver Report mit verknüpften Objekten

Um verschiedene Zeiträume untersuchen zu können, wurde dem Report außerdem eine Zeitachse hinzugefügt, der für beide eingangs beschriebenen Grafiken als Filter fungiert.

Es gibt vielfältige Möglichkeiten, den Report interaktiv auszuwerten, um Wissen aus den Daten zu gewinnen. Beispielhaft halten wir abschließend folgende Erkenntnisse fest:

- Über den gesamten Beobachtungszeitraum gesehen, waren die Rendite und der Umsatz in der Region Ost am größten. Der Umsatz im Norden, Westen und Süden war auf dem gleichen Niveau, aber der prozentuale Gewinn war im Süden am höchsten.
- Die Verteilung des Umsatzes über die Regionen Nord, Ost, Süd und West war war im Laufe der Zeit konstant, wobei die Verteilung des prozentualen Gewinns und dessen absolute Höhe starken Schwankungen unterworfen war.
- Über den gesamten Zeitraum gesehen, war der prozentuale Gewinn in der Sparte *Banking* nur etwa halb so groß wie in der Sparte *Insurance*, obwohl der Umsatz bei *Banking* ca. eineinhalbmal höher war als bei *Insurance*. In den Anfangsjahren des Beobachtungszeitraum war die Verteilung eher ausgeglichen.