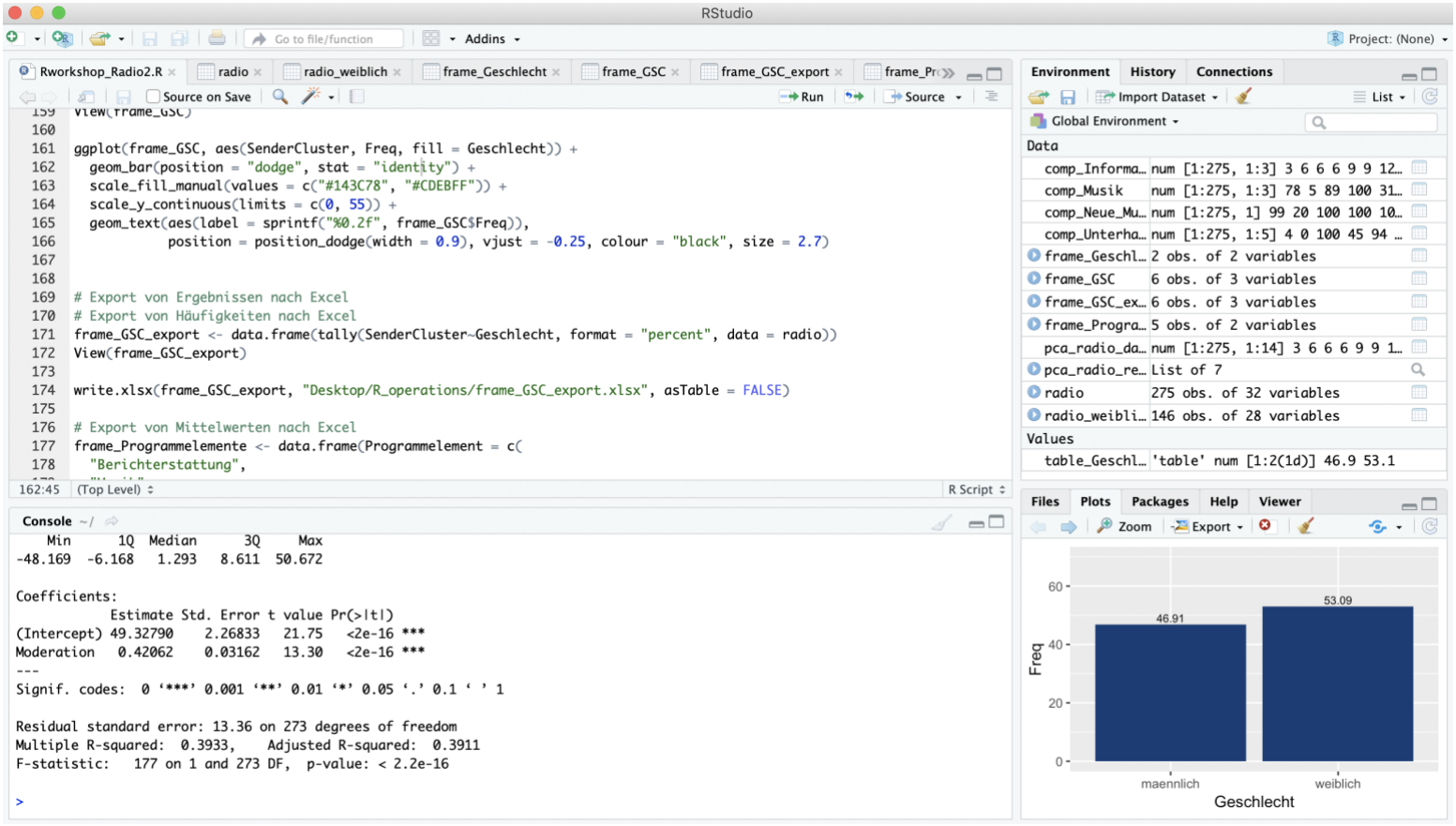
Hendrik Godbersen

Angewandte Statistik & R

Das Buch zum Kompaktkurs mit R-Befehlsblatt für alle gängigen Analyseverfahren, Datensatz zum Üben & Video-Tutorials



Hendrik Godbersen

Angewandte Statistik & R

Inhaltsverzeichnis

1 Einleitung

Dieser Kurs vermittelt das Wissen und die Fähigkeiten:

- die gängigen Verfahren der deskriptiven & inferenzstatistischen Auswertung zu verstehen und anzuwenden.
- mit R zu arbeiten und die für statistische Auswertungen notwendigen Befehle anzuwenden.

Somit sollten Sie nach diesem Kurs, **ein empirisches Projekt mit den „gängigen" statistischen Verfahren in R auszuwerten.**

Um diesen Kurs durchzuführen, benötigen Sie neben diesem Buch die folgenden Unterlagen und Dateien, die Sie auf www.godbersen.online erhalten:

- das **Befehlssammlung mit allen notwendigen R-Befehlen**, um ein empirisches Projekt mit den „gängigen" Verfahren auszuwerten
- den **Datensatz**, mit dem die Übungsaufgaben zu diesem Kurs realisiert werden

Es wird empfohlen, das R-Befehlsblatt auszudrucken, damit die Übungsaufgaben so einfach wie möglich umgesetzt werden können und Sie sich auf die Anwendung der statistischen Verfahren bestmöglich konzentrieren können.

Ferner sind Video-Tutorials Teil des Kurses, die an den entsprechenden Stellen in jedem Kapitel dieses Buchs verlinkt sind. Sie benötigen also eine Internetverbindung, wenn Sie diesen Kurs durchführen. Es wird nachdrücklich dazu geraten, die Video-Tutorials als erstes anzuschauen und den Text in diesem Buch eher als Nachschlagewerk zu betrachten.

Hier gelangen Sie zum **Video-Tutorial** des ersten Kapitels:

Kern des Kurses sind die statistischen Verfahren in den K**apiteln fünf und sechs**. Jedes einzelne Unter-Kapitel behandelt ein Auswertungsverfahren und gliedert sich jeweils in drei Abschnitte:

1) Erklärung der statistischen Verfahren
2) Übungsaufgabe mit einem Datensatz aus einem Forschungsprojekt
3) Umsetzung der statischen Analyse in R

Für den größt- und schnellstmöglichen Lernerfolg ist die **Herangehensweise** in diesem Kurs wichtig. Im Kern sollten Sie sich die Fragen stellen, **wann wende ich welches Verfahren an und wie interpretiere ich die Ergebnisse**. Die Frage, wie programmiere ich mit R, sollte nicht im Vordergrund stehen. Die R-Programmierung werden Sie „nebenbei" lernen.

Darüber hinaus ist es wichtig, dass Sie den **Kurs Schritt für Schritt durchgehen**. Lassen Sie kein Kapitel aus und gehen Sie erst zum nächsten Kapitel über, wenn Sie das vorherige auch verstanden haben. Dies ist wichtig, da die einzelnen Kapitel aufeinander aufbauen. Wenn Sie den Kurs einmal beendet haben, können Sie das Buch, die Video-Tutorials und die Übungsaufgaben als Nachschlagewerk nutzen, z.B. wenn Sie ein bestimmtes Auswertungsverfahren in einem konkreten empirischen Projekt anwenden möchten.

Mit dieser Herangehensweise werden Sie (abhängig von Ihrer Lerngeschwindigkeit etwas schneller oder langsamer) **innerhalb von sieben Stunden solide Kenntnisse und Fähigkeiten in der anwendungsorientierten empirischen Forschung** erhalten. Ziel ist es, dass Sie ein **anspruchsvolles empirisches Projekt mit den gängigen statistischen Verfahren in R analysieren** können.

2 Installation & Grundstruktur von R Studio

In diesem Kapitel lernen Sie, wie Sie R, R Studio und alle notwendigen Zusatzpakete für statistische Analysen installieren. Ferner wird im zweiten Teil ein Überblick über den Aufbau von R Studio, der Benutzeroberfläche für Auswertungen gegeben.

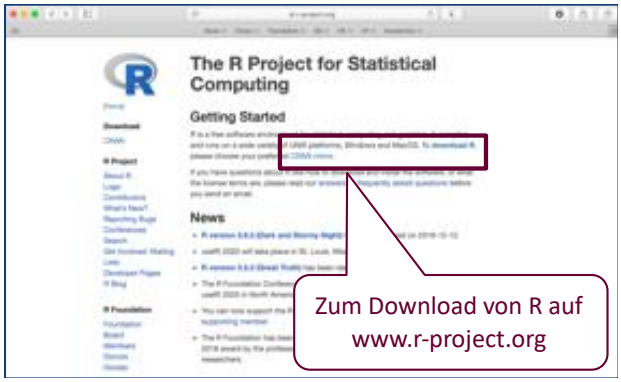
2.1 Installation von R, R Studio & relevanten Paketen

Um Analysen mit R vornehmen zu können, benötigen Sie R, R Studio und einige Zusatzpakete. Im folgenden **Video-Tutorial** sind die notwendigen Installationsschritte beschrieben:

Im Folgenden sind die drei Installationsschritte noch einmal beschrieben:

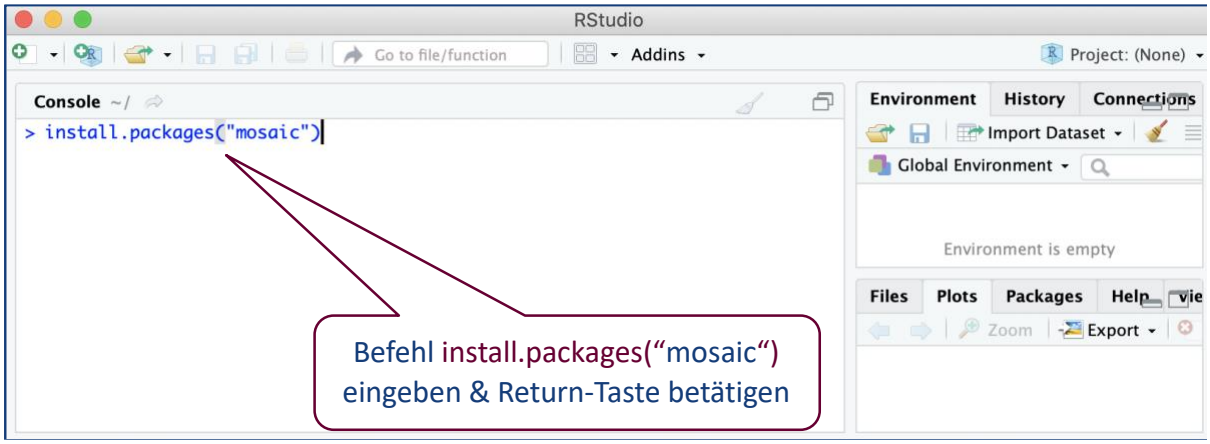
1) **Download & Installation von R** (R ist das Basisprogramm)
    1) Gehen sie zu: https://www.r-project.org/
    2) Wählen Sie einen Mirror (Quelle zum Herunterladen) zum Download aus
        - Achten Sie darauf, dass Sie R passend zu Ihrem Betriebssystem wählen
    3) Laden Sie R herunter
    4) Installieren Sie R auf Ihrem Computer

In der folgenden Abbildung sehen Sie die Einstiegsseite www.r-project.org, über die Sie zum Download von R gelangen:



2) **Download & Installation von R Studio** (R Studio ist eine Entwicklungsumgebung für R und erleichtert die Bedienung)
    1) Gehen Sie zu: https://rstudio.com/products/rstudio/download/
    2) Laden Sie R Studio herunter
    3) Installieren Sie R auf Ihrem Computer
3) **Installieren Sie die Zusatzpakete in R Studio**
    1) Stellen Sie sicher, dass Ihr Computer mit dem Internet verbunden ist
    2) Geben Sie in die Console folgenden Befehl ein: install.packages(“mosaic“)
    3) Betätigen Sie die Return-Taste
    4) Wiederholen Sie dieses Vorgehen für alle relevanten Pakete

In der folgenden Abbildung ist dargestellt, wo Sie in R Studio den Befehl install.packages(““) eingeben müssen, um die notwendigen Zusatzpakete zu installieren:



Die folgenden Pakete benötigen Sie und müssen diese mit dem Befehl install.packages(““) über die Console installieren:

- mosaic
- psych
- psy
- ggplot2
- openxlsx
- readxl

2.2 Grundstruktur von R Studio

In diesem Kapitel geht es um die Grundstruktur von R Studio. Schauen Sie sich dazu das folgende **Video** an:

R Studio ist in vier Bereiche eingeteilt.

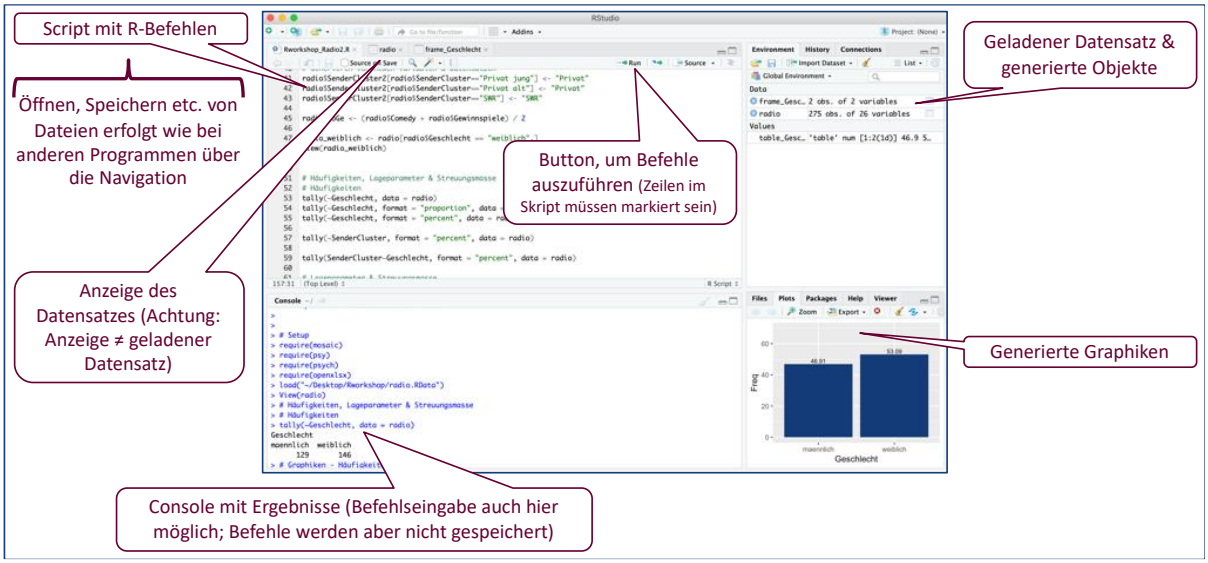
Oben rechts unter Environment werden die geladenen Datensätze angezeigt. (Wie Datensätze geladen werden, wird in Kapitel 4.1 erklärt.)

Oben links wird das Script angezeigt. Im Script werden die Befehle, mit denen ein Datensatz analysiert wird, eingetragen und ausgeführt. Das Erstellen, Öffnen und Speichern von Scripten erfolgt wie bei anderen Anwendungen über das Haupt-Menü. Um einen Befehl auszuführen, muss der Run-Button oberhalb des Scripts betätigt werden; wichtig ist, dass die Zeilen im Script, in denen die auszuführenden Befehle stehen, markiert sind. Im Bereich oben links können Sie sich auch den Datensatz anzeigen lassen; beachten Sie aber, dass der Datensatz nur geladen ist und auch für die Auswertung verwendet werden kann, wenn er oben rechts unter Environment angezeigt wird.

Unten links befindet sich die Console. In dieser werden die Ergebnisse der statistischen Analysen angezeigt. Die Auswertungs-Befehle können auch hier eingegeben und ausgeführt werden; davon wird aber abgeraten, da die Befehle beim Schließen von R verloren gehen. Geben Sie die Befehle deshalb immer im Script (oben links) ein, das Sie speichern, wieder öffnen und dann auch wieder laufen lassen können.

Unten rechts werden unter Plots Graphiken angezeigt, die Sie mit R erstellen. (Das Erstellen von Graphiken wird in Kapitel 7.1 erklärt.)

Die folgende Abbildung gibt einen Überblick über die vier Felder in R Studio:



3 R-Befehle & Übungsdaten

In diesem Kapitel werden die R-Befehlssammlung, das Sie auf www.godbersen.online herunterladen laden können, und die grundsätzliche Befehlsstruktur in R und insbesondere dem Zusatzpaket mosaic, welches das wesentliche Paket für die statistischen Auswertungen ist, erklärt. Ferner wird eine Überblick über die Übungsdaten, die Sie auch auf www.godbersen.online herunterladen können, und das dahinterliegende Forschungsprojekt gegeben.

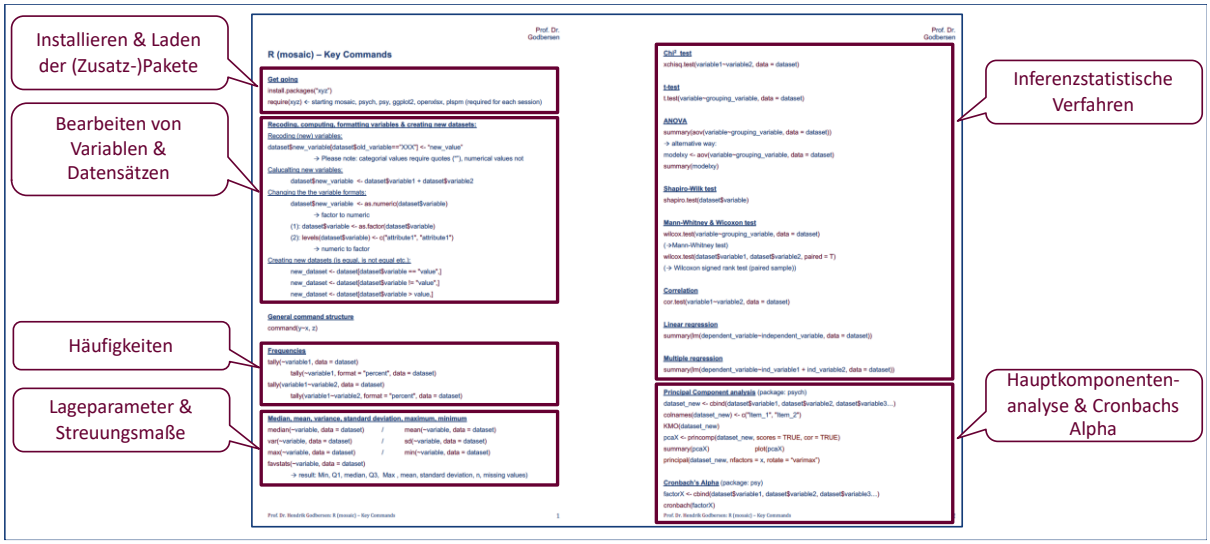
3.1 R-Befehle

Die R-Befehlssammlung besteht aus drei Seiten:

- **Seite 1:** Alle notwendigen Befehle für die Auswertungen von "normalen" Forschungsprojekten
- **Seite 2:** Erstellen von Graphiken und Export von Ergebnissen nach Excel
- **Seite 3:** PLSPM (Partial least path square modelling) – über Standardverfahren hinausgehend und nicht relevant in diesem Kurs

Eine detailliertere Erklärung der R-Befehlssammlung finden Sie im folgenden **Video-Tutorial**:

Die wichtigste Seite (für diesen Kurs) ist die erste, da auf dieser Seite die R-Befehle aller statistischen Standard-Verfahren zu finden sind. Die folgende Abbildung gibt einen Überblick über die erste Seite der R-Befehlssammlung:



Das oben abgebildete **erstes Befehlsblatt gliedert sich in sechs Bereiche**. Oben links finden sich unter „Get going“ die Befehle, die benötigt werden, um Zusatzpakete zu installieren und zu laden. Die Installation muss nur einmal vorgenommen werden. **Das Laden der Pakete (require(mosic) etc.) muss jedes Mal, wenn R Studio gestartet wird, wieder vorgenommen werden.**

Im nächsten Abschnitt darunter („Recoding, computing, formatting variables & creating new datasets“) stehen die Befehle, um die Codes (Merkmalsausprägungen) von Variablen zu verändern, neue Variablen zu berechnen, das Format von Variablen (numerische und nicht numerische Merkmalsausprägungen) anzupassen und aus einem bestehenden Datensatz einen neuen Datensatz zu generieren.

In den nächsten beiden Abschnitten unten rechts auf dem ersten Befehlsblatt finden sich die Befehle zur Auswertung eines Datensatztes nach Häufigkeiten („Frequencies“) und die Bestimmung von Lageparamtern (z.B. Arithmetische Mittelwerte) und Streuungsmaßen (z.B. Standardabweichung) („Median, mean, variance, standard deviation, maximum, mimimum“).

Die rechte Seite des ersten Befehlsblatts ist in zwei Bereich geteilt. Zuerst werden die Befehle für alle Standard-Verfahren der Inferenzstatistik aufgeführt („Chi^2-Test“, „t-Test“ etc.). Daran anschließend werden unten links die Befehle für die Hauptkomponentenanalyse („Principal component cnalysis“) als strukturentdeckendes Verfahren und die Berechnung von

Cronbachs alpha, die im Zusammenhang mit der Hauptkomponentenanalyse durchgeführt wird („Cronbach's alpha"), dargestellt.

Die **Grundstruktur der Befehle** im R-Zusatzpaket mosaic, welches das hauptsächliche Auswertungs-Tool ist, stellt sich, wie folgt, dar:

- command(y~x, z)

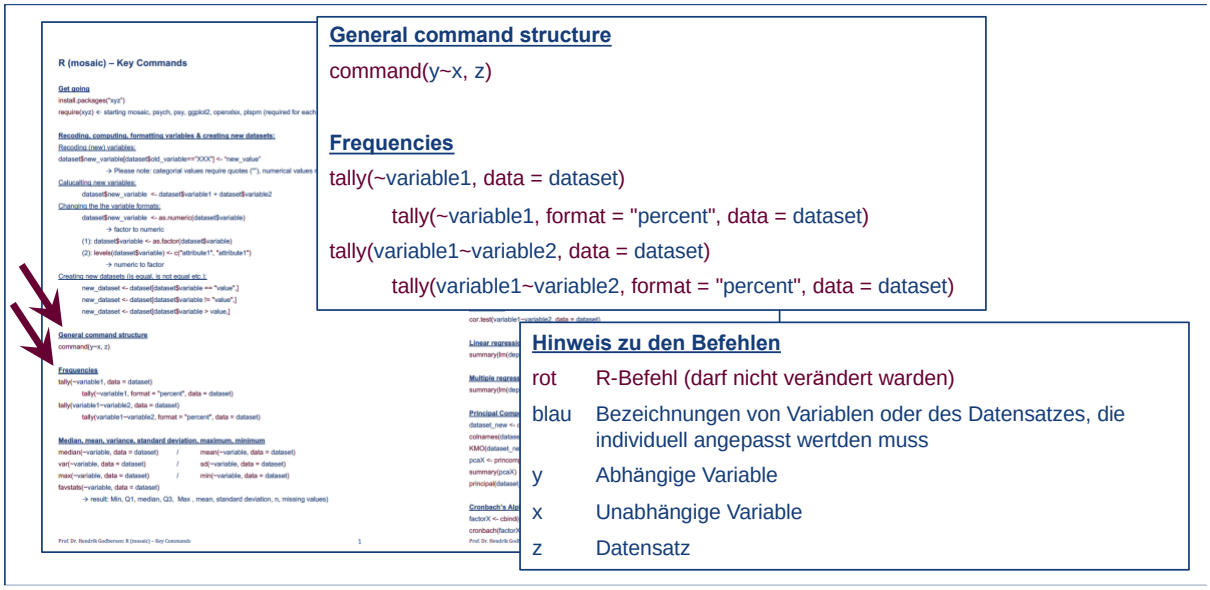
Als erstes wird der Befehl für das Auswertungsverfahren (Häufigkeit, arithmetischer Mittelwert, Chi^2-Test, t-Test etc.) angegeben. Dann folgt eine öffnende Klammer. Die abhängige Variable wird angegeben. Nach dem Schlangenzeichen folgt die unabhängige Variable. Es kommt ein Komma und der Datensatz, aus dem R sich die Variablen ziehen soll, wird eingetragen. Der Befehl wird mit einer abschließenden Klammer beendet. Wenn es keine zwei Variablen oder keine abhängige Variable gibt, entfällt das y und der Befehl beginnt mit dem Schlagenzeichen.

Ein Beispiel soll diese Struktur verdeutlichen. Es soll herausgefunden werden, wieviele Frauen und Männern, die in der Variable „Geschlecht" erfasst wurden, im Datensatz „radio" gegeben sind. Dazu wäre folgender Befehl notwendig (eine detailliertere Einführung in die Auswertung mit Häufigkeiten finden Sie in Kapitel 5.2):

- tally(~Geschlecht, data = radio)

Die Farben in der Befehlssammlung sind für die Anwendung von Bedeutung. Alle Buchstaben und Zeichen, die rot dargestellt sind, werden als fixe Befehle von R so gefordert und dürfen nicht verändert werden. Die blauen Wörter sind Variablen oder der auszuwertende Datensatz, die je nach ihrer individuellen Bezeichnung angepasst werden müssen.

Die soeben beschriebene Befehlsstruktur ist noch einmal in der folgenden Abbildung dargestellt:



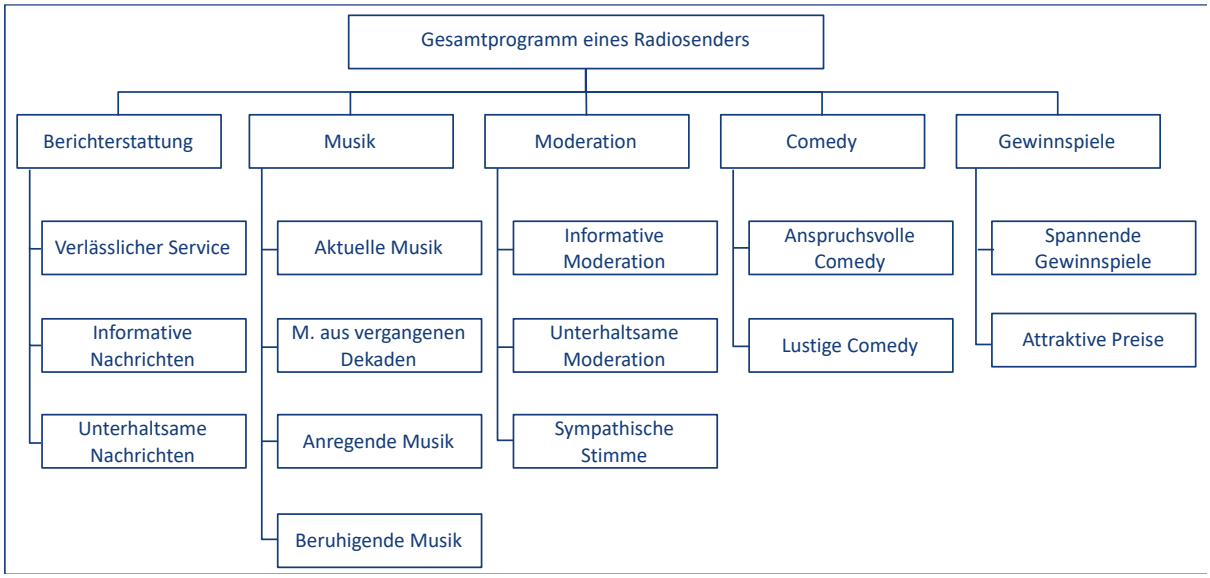
3.2 Übungsdaten

Im Rahmen dieses Kurses müssen Sie alle gängigen Auswertungsverfahren in Übungsaufgaben anwenden und so selbst eine statistische Auswertung vornehmen. Diese Auswertung erfolgt mit einem vereinfachten Datensatz aus einem Forschungsprojekt, das in folgendem Artikel veröffentlicht wurde:

- Godbersen, H. (2019): Hörererwartungen, Programmqualität und Optimierungspotenzial von musikbasierten Formatradios – Eine empirische Analyse mit der Means-End Theory of Complex Cognitive Structures, in: transfer – Zeitschrift für Kommunikation und Markenmanagement, Jg. 65, Nr. 3, S. 12-21

Zur näheren Erläuterung des Datensatzes und der dahinterliegenden inhaltlichen Überlegungen schauen Sie sich das folgende **Video-Tutorial** an:

Das o.a. Forschungsprojekt behandelt die Beurteilung von Formatradiosendern. Die Grundstruktur des Beurteilungsmodells ist in der folgenden Abbildung dargestellt:



Das Modell besteht aus drei Ebenen. An der Spitze steht die Qualität des Gesamtprogramms aus Hörersicht. Auf der darauffolgenden Ebene stehen die Programmelemente Berichterstattung, Musik, Moderation, Comedy und Gewinnspiele. Auf der untersten und konkretesten Ebene sind die Programmeigenschaften (z.B. aktuelle Musik, Musik aus vergangenen Dekaden, anregende und beruhigende Musik) zu finden. Die Variablen auf den genannten drei Ebenen wurden mit Continuous Rating Scales von 0 (nicht gut) bis 100 (sehr gut) erfasst. Daneben wurden weitere Variablen erfasst:

- Stammsender: in den letzten zwei Wochen am häufigsten gehörter Sender
- SenderCluster: neugebildete Variable mit den Merkmalsausprägungen Privat jung, Privat alt und SWR
- Geschlecht: mit den Merkmalsausprägungen männlich und weiblich
- Alter: angegebenen in Jahren
- Altersgruppen: zwei Gruppen mit den Merkmalsausprägungen jung und alt
- Affinität zu aktueller Musik (Variablen-Bezeichnung: Aktuelle_Musik_A): zwei Gruppen mit den Merkmalsausprägungen „eher affin zu aktueller Musik“ und „eher nicht affin zu aktueller Musik“

Eine Gesamtübersicht zu den Variablen des Datensatzes finden Sie in der folgenden Abbildung, wobei die Variablen-Bezeichnungen rot geschrieben sind:

- Stammsender
  - Messung: in den letzten 2 Wochen am häufigster gehörter Sender
- SenderCluster
  - Messung: neu gebildete Variable mit drei Ausprägungen
  - Merkmalsausprägungen: SWR / Privat jung / Privat alt
- Programmeigenschaften
  - Messung: Continuous Rating Scale von 0 (nicht gut) bis 100 (sehr gut)
  - 14 Items
    - Verlaesslicher_Service
    - Informative_Nachrichten
    - Unterhaltsame_Nachrichten
    - Aktuelle_Musik
    - Alte_Musik
    - Anregende_Musik
    - Beruhigende_Musik
    - Informative_Moderation
    - Unterhaltsame_Moderation
    - Sympathische_Stimme
    - Anspruchsvolle_Comedy
    - Lustige_Comedy
    - Spannende_Gewinnspiele
    - Attraktive_Preise

Hinweis: Variablennamen in rot

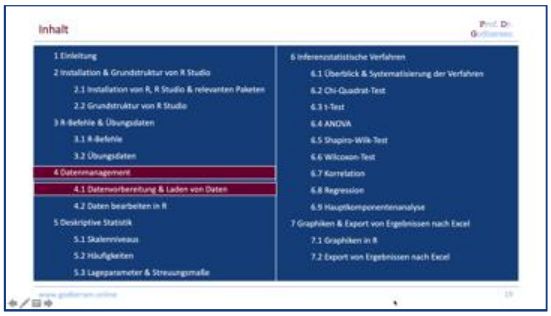
- Programmelemente
  - Messung: Continuous Rating Scale von 0 (nicht gut) bis 100 (sehr gut)
  - 5 Items
    - Berichterstattung
    - Musik
    - Moderation
    - Comedy
    - Gewinnspiele
- Gesamtprogramm
  - Messung: Continuous Rating Scale von 0 (nicht gut) bis 100 (sehr gut)
- Geschlecht
  - Messung: Auswahlfrage
- Alter
  - Messung: offene Altersangabe
- Altersgruppen (alt vs. jung, 2 Gruppen)
- Aktuelle_ Musik_A (Affinität, 2 Gruppen)

4 Datenmanagement

In diesem Kapitel lernen Sie, wie Sie Daten für die Analyse mit R vor- und aufbereiten und wie Sie diesen Datensatz dann in R laden und für weitere R-Sitzungen bereithalten. Daran anschließend wird die Bearbeitung von Daten in R Studio erklärt. In diesem Kapitel beginnen Sie mit der Bearbeitung der Übungsaufgaben und werden selbst mit R arbeiten.

4.1 Datenvorbereitung & Laden von Daten

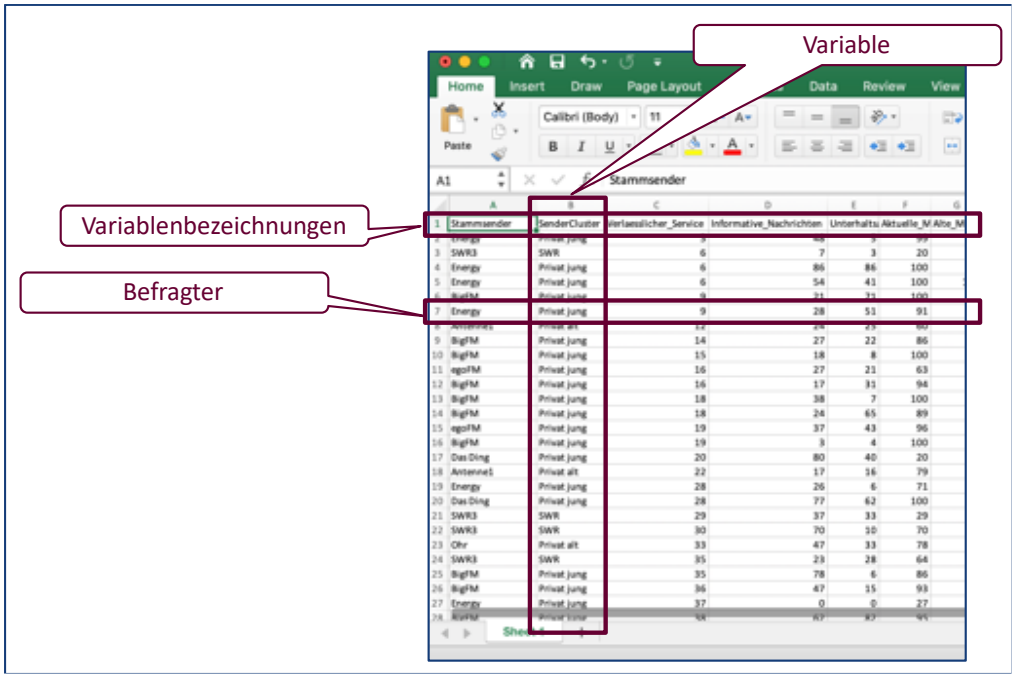
Die Auswertung in einem Forschungsprojekt kann nur so gut sein wie die dahinterliegenden Daten. Dementsprechend ist es wichtig, dass Sie bei jedem Forschungsprojekt die Daten „in Ordnung halten/bringen“ und dann in R laden. Im folgenden **Video-Tutorial** werden die Datenvorbereitung und das Laden des Datensatzes in R erklärt und demonstriert:



Im ersten Schritt geht es um die Datenvorbereitung. Laden Sie sich die Daten vom Server, auf dem Sie Ihre Online-Befragung durchgeführt haben, im Excel-Format herunter. Andere Formate sind auch möglich, jedoch zeigt die Erfahrung, dass im Excel-Format die wenigsten Probleme auftreten. Sollten Sie einen schriftlichen Fragenbogen eingesetzt haben oder die Daten über ein anderes Medium erhoben haben, überführen Sie diese in eine Excel-Tabelle.

Die Datenstruktur sollte, wie folgt, aussehen und ist auch in der nachfolgenden Abbildung dargestellt:

- Erste Zeile = Variablenbezeichnungen
- 1 Spalte = 1 Variable
- 1 Zeile = 1 Befragter



Eine Empfehlung ist, nur vollständig ausgefüllte Fragebögen in den Datensatz und die spätere Analyse aufzunehmen. Dies vereinfacht den Umgang mit fehlenden Werten, die bei diesem Vorgehen schlichtweg nicht mehr existieren. Die Entscheidung darüber hängt jedoch vom Forschungsprojekt ab.

Voraussetzung dafür, dass Sie die Daten **mit R auswerten** können, ist hinsichtlich:

- **Variablenbezeichnungen:** keine Leerzeichen & keine Umlaute
  - Empfehlung: kurze & bedeutungsvolle Bezeichnungen
- **Merkmalsausprägungen:**
  - Metrische Variablen (z.B. Alter) als numerischen Wert
  - Kategoriale Variablen (z.B. Geschlecht) als Text

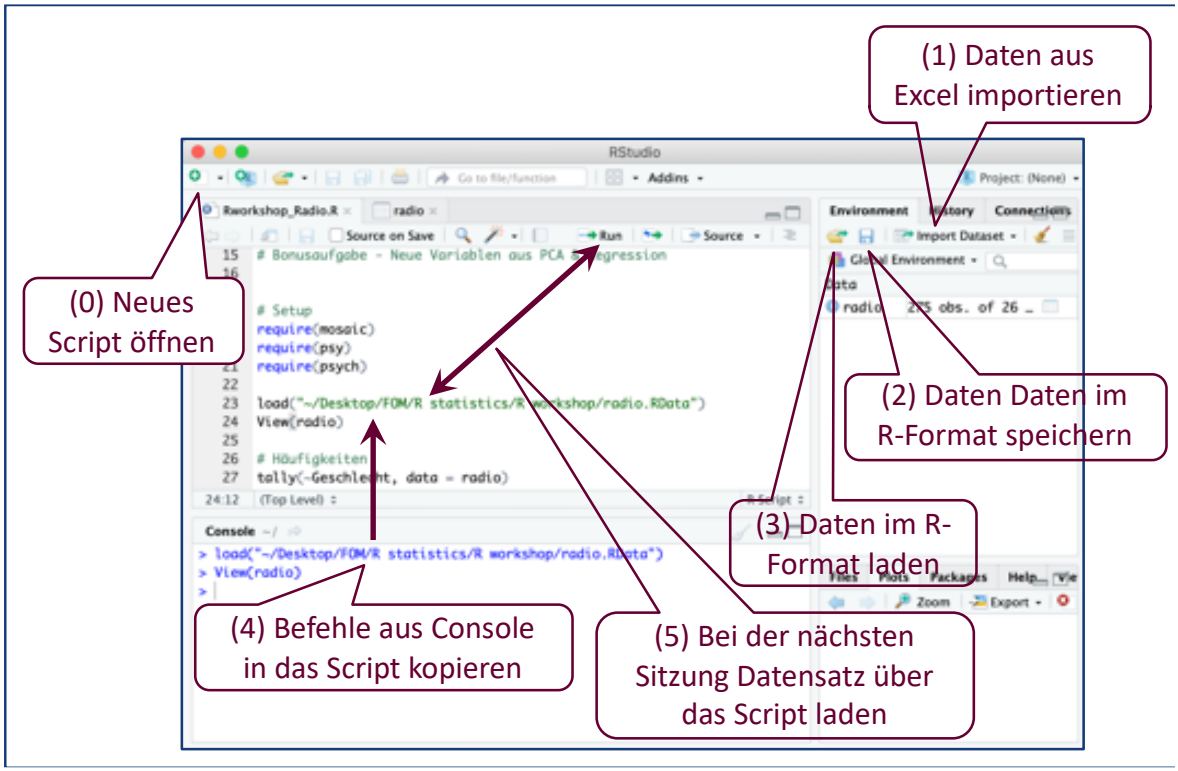
Im zweiten Schritt müssen Sie die **Daten in R Studio laden** und auch für Ihre nächsten R-Sitzungen bereitstellen. Dazu sind die folgenden fünf Schritte notwendig, die Sie im Detail und ohne Ausnahme anwenden sollten/müssen:

0) Öffnen Sie ein neues Script über das Hauptmenü

1) **Vorbereitung (<u>wichtig</u>):**
   - Speichern Sie die Daten in einem Ordner auf Ihrem Computer, der extra für das Projekt angelegt ist

2) **Daten aus Excel importieren:**
   - Unter „Environment“ Daten über „Import Dataset“ Excel-Daten auswählen
3) **Daten im R-Format speichern:**
   - Unter Environment über Speichersymbol Daten im Projektordner speichern
4) **Daten im R-Format laden:**
   - Unter „Environment“ Daten über Öffnungssymbol R-Daten auswählen
   - Unter „Environment“ auf den Datensatz klicken
   - In der Console generierte Befehle in das Script kopieren
5) **Bei der nächsten Anwendung können die Daten über das Script gestartet werden**
   - Zeile mit dem load-Befehl markieren und Run-Button betätigen

Die fünf oben beschriebenen Schritte sind im Folgenden noch einmal graphisch dargestellt:



4.2 Daten bearbeiten in R

In vielen Forschungsprojekten ist es notwendig den erhobenen Datensatz zu bearbeiten. Es kann geboten sein, Variablen zu recodieren, neue Variablen zu berechnen oder einen neuen Datensatz zu erstellen. Im Folgenden finden Sie für die genannten drei Fälle Beispiele:

- Recodieren von Variablen:
  - Inverse Skalen, z.B. 1 zu 6, 2 zu 5, 3 zu 4…
  - Zusammenfassen von Werten, z.B. Einwohner von Bayern und Baden-Württemberg (erhobene Variable) zu Süddeutschland (neu zu bildende Variable)
  - Ändern von metrischen zu kategorialen Werten (siehe Datenvorbereitung für R)
  - …
- Berechnen neuer Variablen:
  - Bildung eines neuen („übergeordneten“) Konstrukts aus mehreren Variablen, z.B. Berechnung einer Gesamtbewertung aus mehreren Einzelbewertungen über das arithmetische Mittel
  - Berechnung einer neuen Variablen, die sich aus dem Verhältnis mehrerer Variablen bestimmt, z.B. Berechnung der Rentabilität aus den Variablen Gewinn und Umsatz
  - …
- Erstellen neuer Datensätze:
  - Ausschluss von Befragten mit bestimmten Merkmalsausprägungen, z.B. zu alte oder zu junge Befragte
  - Analyse nur eines Teils der Befragten, z.B. ausschließliche Analyse der Antworten von Frauen
  - …

Schauen Sie sich nun das **Video-Tutorial** an, in dem die Bearbeitung von Daten in R erklärt wird und zu allen drei oben genannten Datenbearbeitungen eine Übungsaufgabe gestellt wird:

Wie im Video-Tutorial dargestellt, finden Sie im Folgenden noch einmal die Aufgaben und Lösungen zur Bearbeitung von Daten.

Aufgaben:

- Generieren Sie aus SenderCluster eine neue Variable (SenderCluster2): “Privat jung“ & „Privat alt“ zu „Privat“ und „SWR“ zu „SWR“.
- Berechnen Sie aus den Variablen „Comedy“ und „Gewinnspiele“ in einer neuen Variablen (CoGe) den Mittelwert.
- Generieren Sie einen neuen Datensatz, der nur aus Frauen besteht.

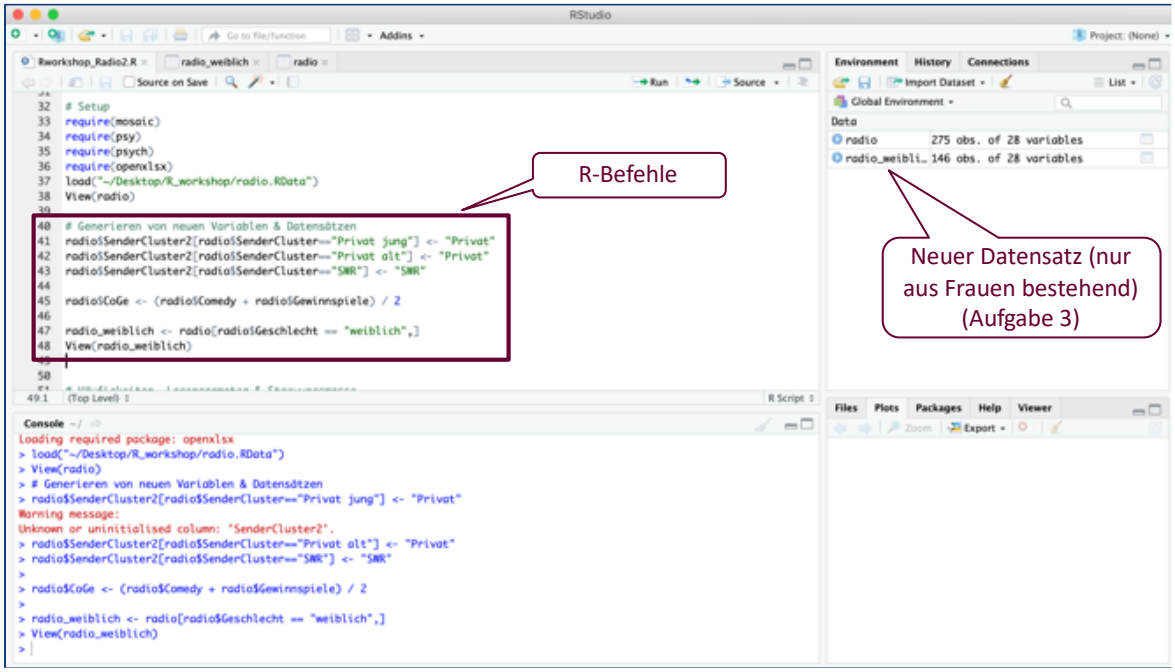
Benötigte Befehle:

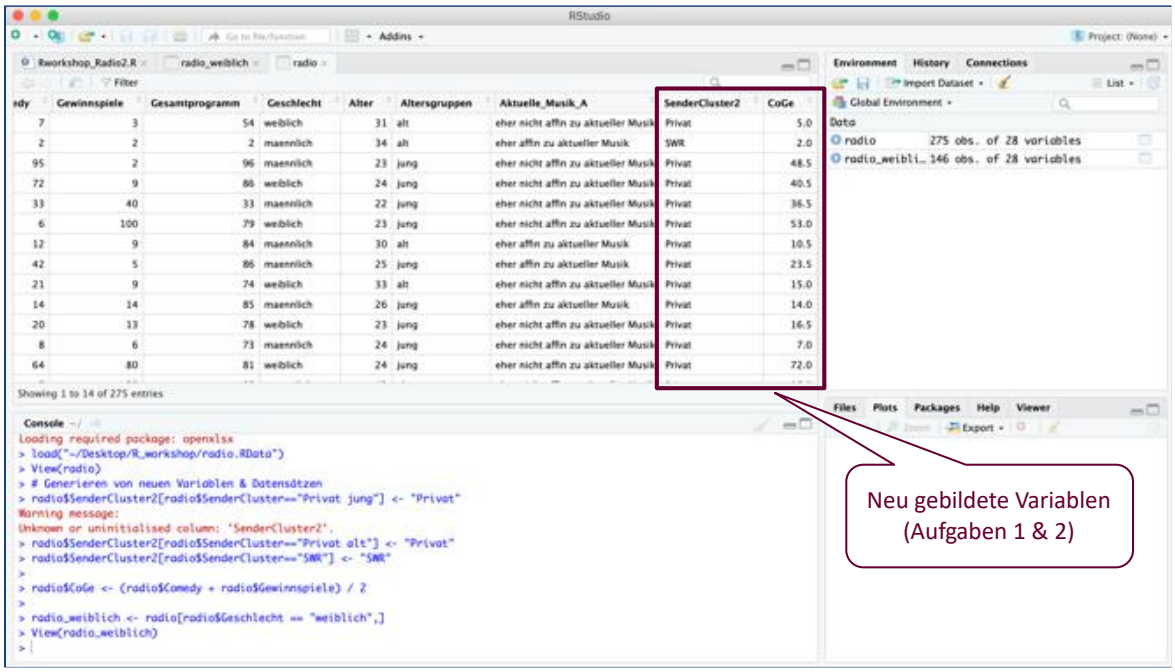
```
Recoding (new) variables:
dataset$new_variable[dataset$old_variable==”XXX”] <- “new_value”
        → Please note: categorial values require quotes (“”), numerical values not
Calculating new variables:
    dataset$new_variable  <- dataset$variable1 + dataset$variable2
Changing the the variable formats:
    …
Creating new datasets (is equal, is not equal etc.):
    new_dataset <- dataset[dataset$variable == "value",]
    new_dataset <- dataset[dataset$variable != "value",]
    new_dataset <- dataset[dataset$variable > value,]
```

Relevante Variablen:

- SenderCluster (Aufgabe 1)
- Comedy und Gewinnspiele (Aufgabe 2)
- Geschlecht (Aufgabe 3)

Lösung in R:


R-Befehle
Neuer Datensatz (nur aus Frauen bestehend) (Aufgabe 3)

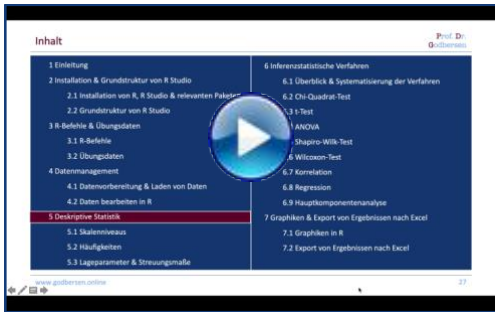

Neu gebildete Variablen (Aufgaben 1 & 2)

5 Deskriptive Statistik

Mit dem fünften Kapitel beginnt der Kern des Kurses, die Vermittlung der statistischen Verfahren und deren Anwendung in R. Als Voraussetzung dafür werden im ersten Unterkapitel die Skalenniveaus erklärt. Darauf aufbauend werden Häufigkeiten sowie Lageparameter und Streuungsmaße behandelt. Auch wenn Sie schon mit diesen deskriptiven statistischen Verfahren vertraut sind, wird nachdrücklich empfohlen, diese Kapitel sorgfältig durchzugehen, da hier die Arbeitsweise in R deutlich wird. Dementsprechend finden Sie in diesem Kapitel auch mehrere Übungsaufgaben, so dass Sie nach Abschluss des Kapitels mit der Benutzung von R vertraut sein sollten.

Wichtiger Hinweis: Stellen Sie sicher, dass Sie vor der Auswertung **bei jeder neuen Sitzung** die benötigten **Zusatzpakete geladen** haben über den Befehl **require(mosaic)** (für das Laden der anderen Pakete ersetzen Sie mosiac mit dem entsprechenden Paket).

Schauen Sie sich auch das folgende **Video-Tutorial** an:



5.1 Skalenniveaus

Für die Anwendung von statistischen Verfahren ist es wichtig, welches Skalenniveau die zu analysierenden Variablen haben. Schauen Sie sich dazu das folgende **Video** an:

Für Anwendung von Häufigkeiten und Lageparametern sowie auch für die in Kapitel sechs erklärten inferenzstatistischen Verfahren ist es notwendig, zwischen zwei Arten von Variablen zu unterscheiden:

- **Kategoriale Variablen** mit nominalem und ordinalem Skalenniveau
- **Metrische Variablen** mit intervallskaliertem und verhältnisskaliertem Niveau

Eine Aufteilung der Skalenniveaus mit mathematischen Eigenschaften, Messwertcharakteristik, Beispiel und möglichen Lageparametern finden Sie in der folgenden Tabelle:

Skalenniveau		Math. Eigenschaften	Beschreibung der Messwertcharakteristika	Beispiel	Lageparameter
Kategorial	Nominal	=/≠	Die Messwerte sind identisch oder nicht identisch.	Geschlecht	• Modus
	Ordinal	=/≠ ; </>	Die Messwerte sind größer, kleiner oder identisch.	Olympische Ränge	• Modus • Median
Metrisch	Intervall	=/≠ ; </> ; −; +	Die Distanz zwischen den Messwerten kann angegeben werden.	Temperatur	• Modus • Median • Arithm. Mittel
	Verhältnis	=/≠ ; </> ; +/− ; */÷	Die Distanz zwischen den Messwerten und deren Verhältnis kann angegeben werden.	Körpergröße	

5.2 Häufigkeiten

In den meisten Auswertungen sind drei Arten von Häufigkeiten relevant, wobei i.d.R. die relative Häufigkeit am wichtigsten ist:

- **Absolute Häufigkeit**: gibt an, wie häufig ein Wert auftritt
- **Relative Häufigkeit**: gibt das Verhältnis eines untersuchten Merkmalsträgers zu einem Wert oder einer Klasse von Werten an
- **Kreuztabelle**: stellt die gemeinsame Verteilung zweier Variablen dar

Im folgenden **Video-Tutorial** werden die Häufigkeiten, ihre statistische Verwendung und die Umsetzung in R, inklusive Übungsaufgaben, erläutert:

Aufgaben 1:

- Wie viele Männer und Frauen gibt es in der Stichprobe?
- Wie hoch ist der Anteil der Männer und Frauen in %?
- Wie hoch ist der Anteil der SenderCluster in %?

Benötigte Befehle 1:

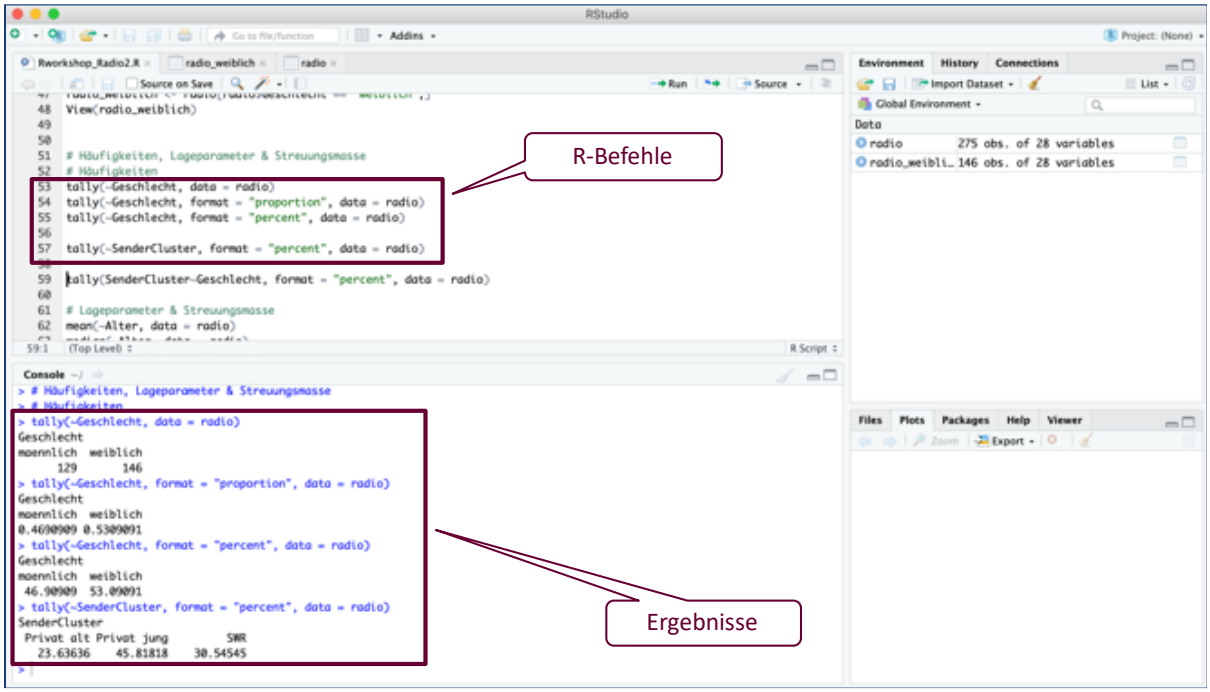
Frequencies

```
tally(~variable1, data = dataset)
        tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
        tally(variable1~variable2, format = "percent", data = dataset)
```

Relevante Variablen 1:

- SenderCluster
- Geschlecht

Lösung in R 1:



Aufgaben 2:

- Wie verteilen sich Männer und Frauen auf die SenderCluster (%)?

Benötigte Befehle 2:

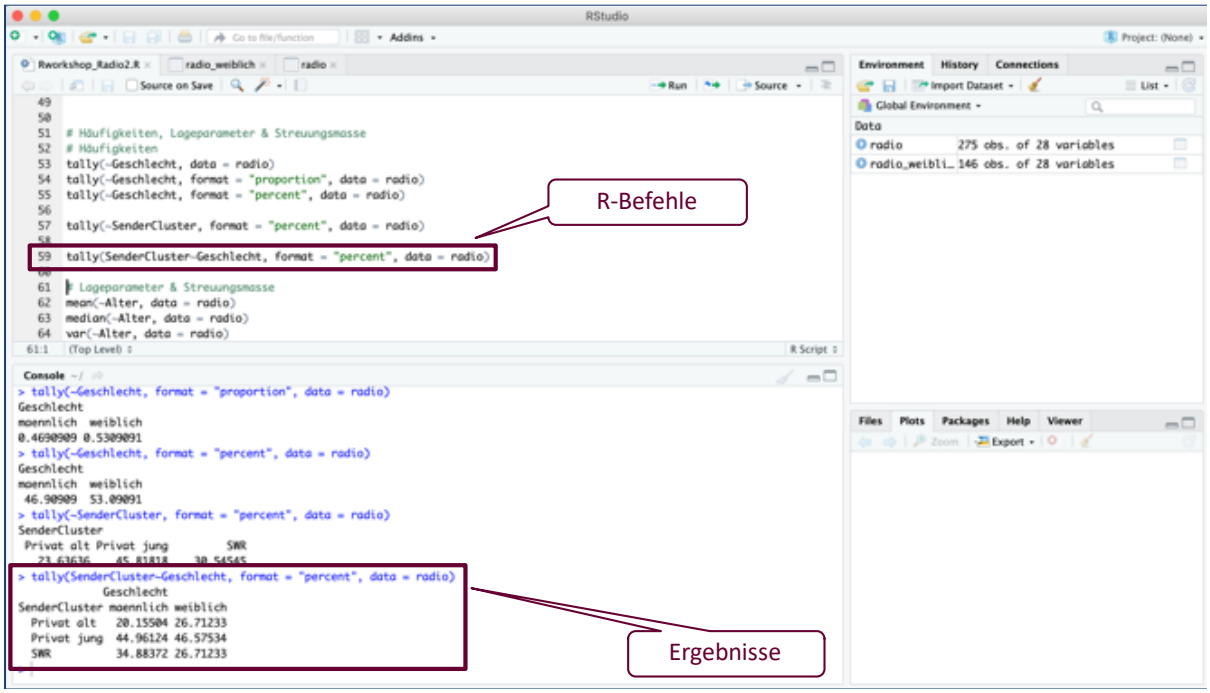
Frequencies

```
tally(~variable1, data = dataset)
        tally(~variable1, format = "percent", data = dataset)
tally(variable1~variable2, data = dataset)
        tally(variable1~variable2, format = "percent", data = dataset)
```

Relevante Variablen 2:

- SenderCluster
- Geschlecht

Lösung in R 2:



5.3 Lageparameter & Streuungsmaße

Im Folgenden sind die **Lageparameter** definiert:

- **Arithmetisches Mittel**: Division der Summe aller Werte durch die Anzahl der Werte; metrisches Skalenniveau notwendig
- **Median**: Wert, der die Grenze zwischen oberer und unterer Hälfte markiert; mindestens ordinales Skalenniveau
- **Modalwert**: Häufigster Wert einer Häufigkeitsverteilung; mindestens nominales Skalenniveau

Neben den Lageparametern sind die **Streuungsmaße** für die Interpretation von Daten relevant. Die beiden zentralen Streuungsmaße Varianz und Standardabweichung sind, wie folgt, definiert:

- **Streuungsparameter**
  - **Varianz**: Durchschnittliche quadrierte Abweichung der gemessenen Werte vom arithmetischen Mittel (beachte: empirische Varianz → /n; Stichprobenvarianz → /n-1)

- **Standardabweichung**: Durchschnittliche Abweichung der gemessenen Werte vom arithmetischen Mittel

Schauen Sie sich auf jeden Fall das nächste Video-Tutorial an, in dem unter anderem die Lageparameter und die Streuungsmaße detaillierter erklärt werden. Es sei an dieser Stelle darauf verwiesen, dass die Varianz (Streuungsmaß) die mathematische Grundlage für die meisten inferenzstatistischen Verfahren ist. **(„Wenn Sie die Varianz verstanden haben, haben Sie den Kern der (inferenz-)statistischen Analyseverfahren verstanden“**):

Aufgaben 1:

- Was ist das durchschnittliche Alter der Befragten (arith. Mittel)?
- Welchen Wert hat der Median bei der Variable Alter?
- Wie stark streuen die gemessenen Werte der Variable Alter um das arithmetische Mittel (Varianz & Standardabweichung)?

Benötigte Befehle 1:

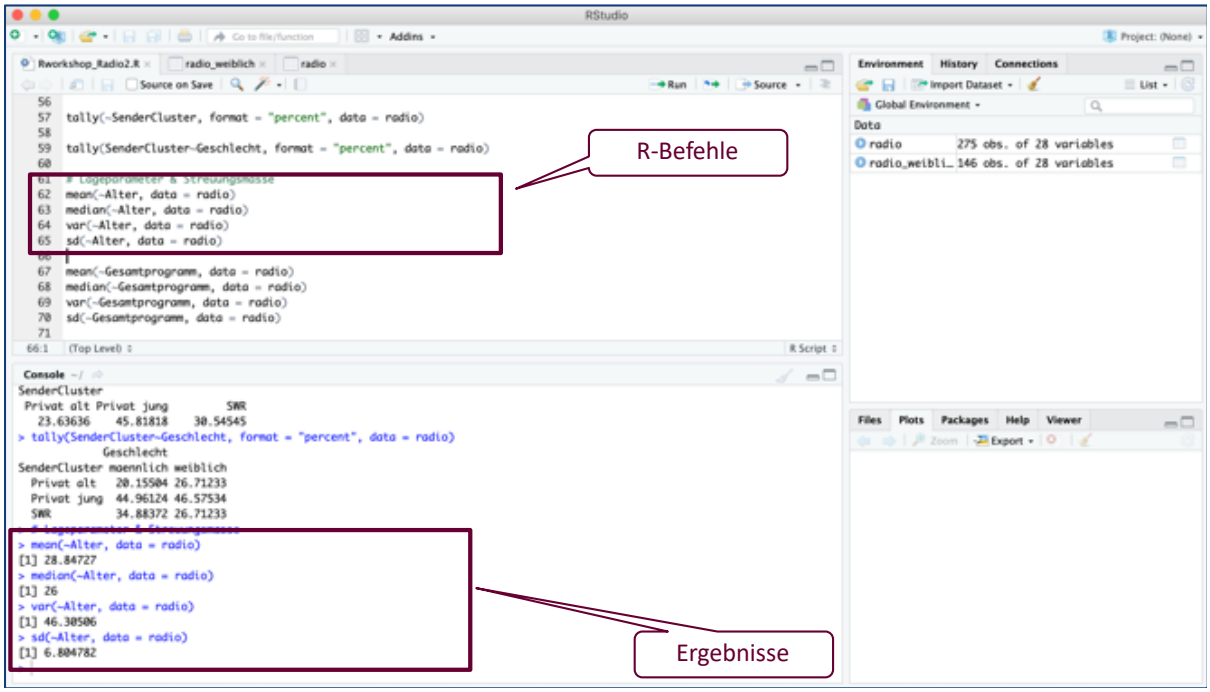
Median, mean, variance, standard deviation, maximum, minimum

```
median(~variable, data = dataset) / mean(~variable, data = dataset)
var(~variable, data = dataset) / sd(~variable, data = dataset)
max(~variable, data = dataset) / min(~variable, data = dataset)
favstats(~variable, data = dataset)
```

Relevante Variablen 1:

- Alter

Lösung in R 1:



Aufgaben 2:

- Wie hoch ist die durchschnittliche Bewertung des Gesamtprogramms (arith. Mittel)?
- Welchen Wert hat der Median bei der Variable Gesamtprogramm?
- Wie stark streuen die gemessenen Werte der Variable Gesamtprogramm um das arithmetische Mittel (Varianz & Standardabweichung)?

Benötigte Befehle 2:

Median, mean, variance, standard deviation, maximum, minimum

median(~variable, data = dataset) / mean(~variable, data = dataset)

var(~variable, data = dataset) / sd(~variable, data = dataset)

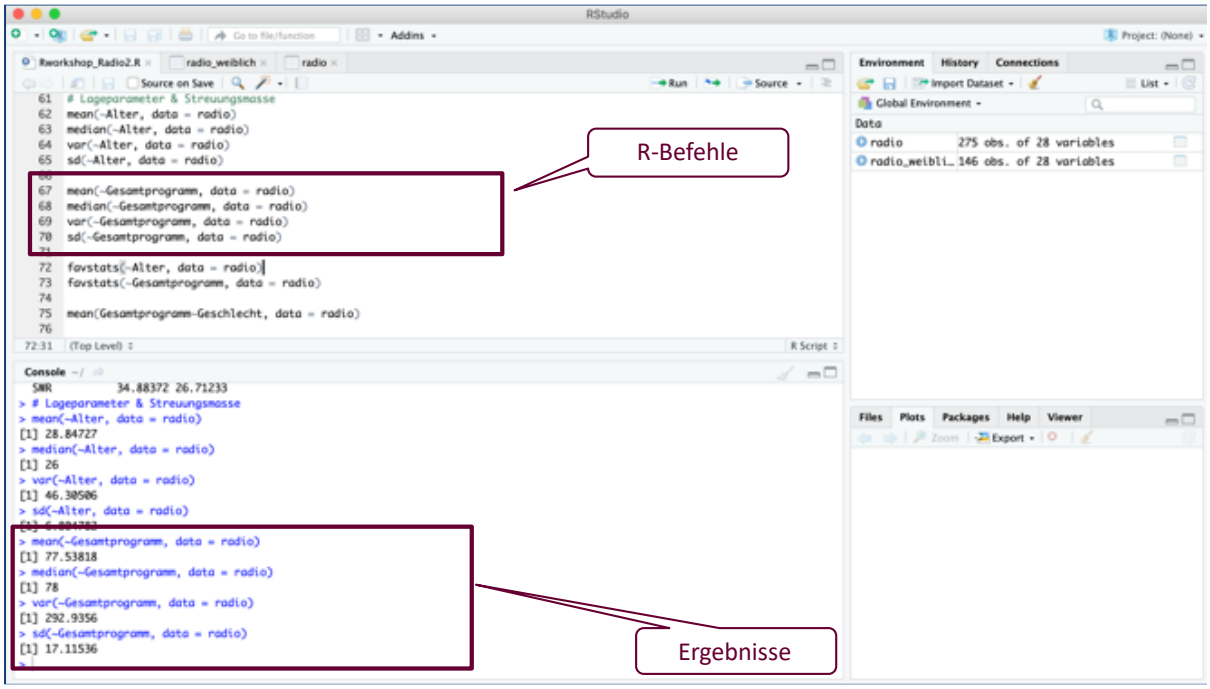
max(~variable, data = dataset) / min(~variable, data = dataset)

favstats(~variable, data = dataset)

Relevante Variablen 2:

- Gesamtprogramm

Lösung in R 2:



Aufgaben 3:

- Bestimmen Sie für die Variable Alter das arithmetische Mittel, den Median, die Standardabweichung etc. mit nur einem R-Befehl.
- Bestimmen Sie für die Variable Gesamtprogramm das arithmetische Mittel, den Median, die Standardabweichung etc. mit nur einem R-Befehl.

Benötigte Befehle 3:

Median, mean, variance, standard deviation, maximum, minimum

median(~variable, data = dataset) / mean(~variable, data = dataset)

var(~variable, data = dataset) / sd(~variable, data = dataset)

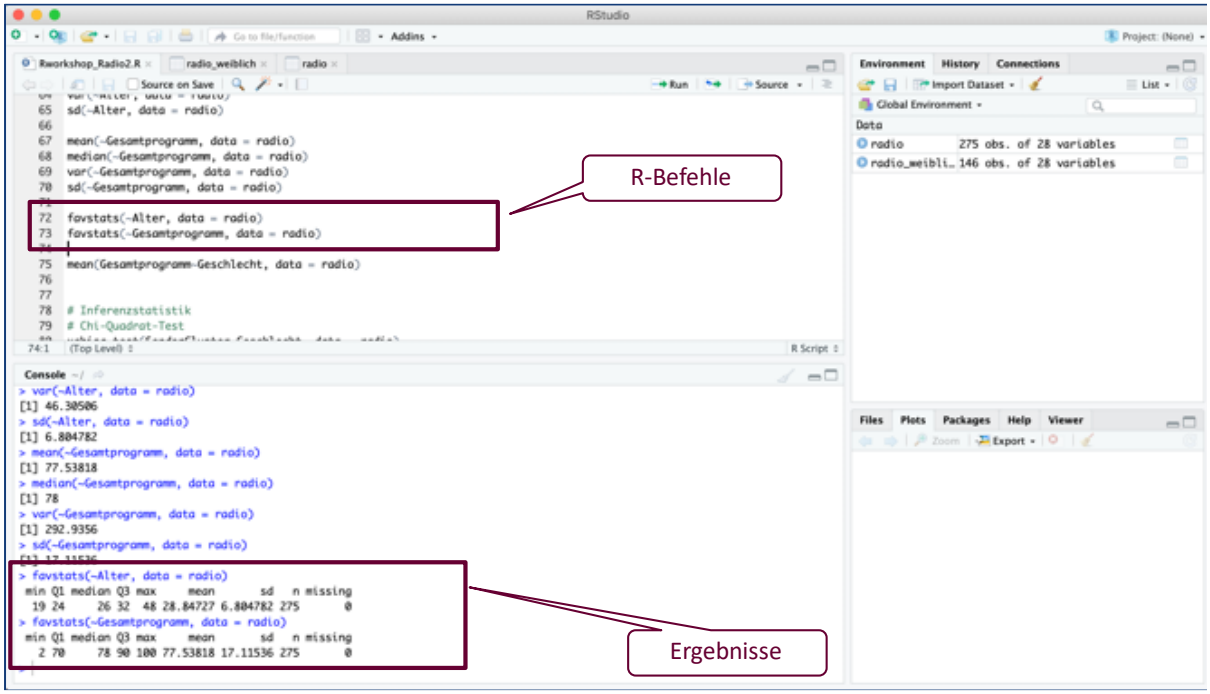
max(~variable, data = dataset) / min(~variable, data = dataset)

favstats(~variable, data = dataset)

Relevante Variablen 3:

- Alter
- Gesamtprogramm

Lösung in R 3:



Aufgaben 4:

- Stellen Sie das durchschnittliche Alter von Männern und Frauen gegenüber (arith. Mittelwerte).
- Stellen Sie die durchschnittliche Bewertung des Gesamtprogramms von Männern und Frauen gegenüber (arith. Mittelwerte).

Benötigte Befehle 4:

Median, mean, variance, standard deviation, maximum, minimum

median(~variable, data = dataset) / mean(~variable, data = dataset)

var(~variable, data = dataset) / sd(~variable, data = dataset)

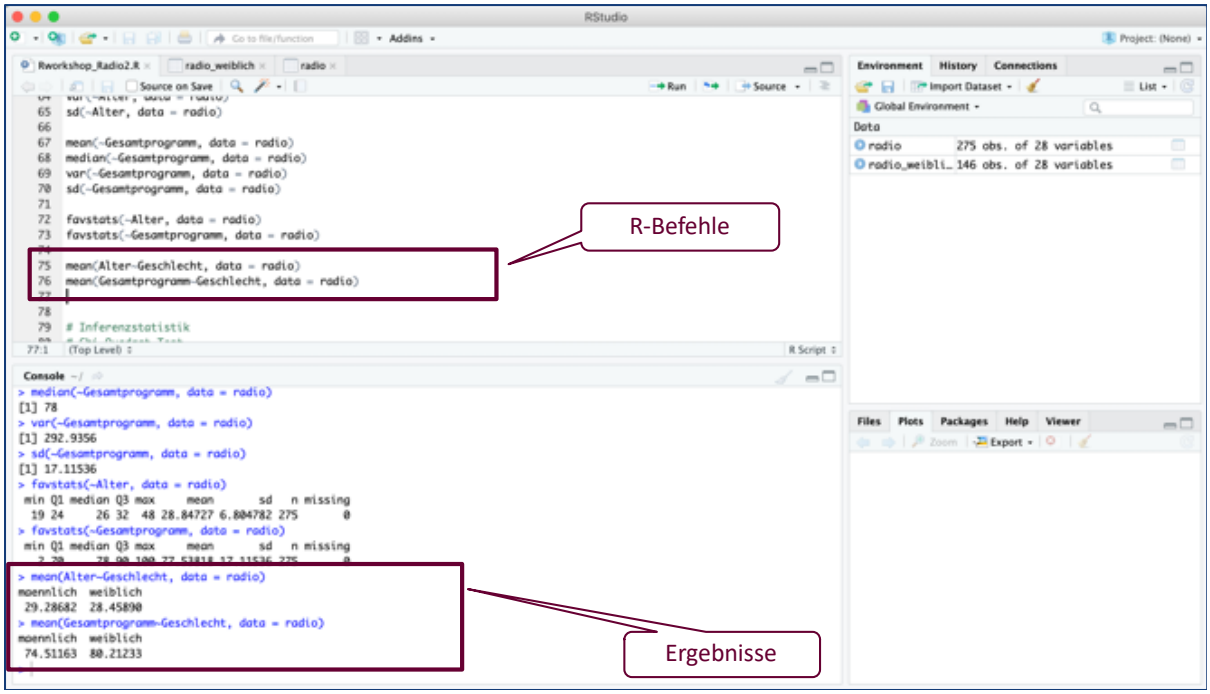
max(~variable, data = dataset) / min(~variable, data = dataset)

favstats(~variable, data = dataset)

Relevante Variablen 4:

- Alter
- Gesamtprogramm
- Geschlecht

Lösung in R 4:


RStudio
R-Befehle
Ergebnisse
Console
Environment
History
Connections
Files
Plots
Packages
Help
Viewer

6 Inferenzstatistische Verfahren

In diesem Kapitel wird zuerst ein Überblick über die Verfahren der strukturprüfenden Analyse (Inferenzstatistik) und der strukturentdeckenden Analyse gegeben, so dass Sie wissen, wann Sie welches Verfahren anwenden müssen. Daran anschließend werden die einzelnen Verfahren in jeweils einem Unterkapitel erklärt, Sie wenden diese in Übungsaufgaben an und lernen, wie Sie die Ergebnisse interpretieren müssen.

6.1 Überblick & Systematisierung der Verfahren

In der Inferenzstatistik wird geprüft, ob ein signifikanter (nicht zufälliger) Unterschied zwischen zwei oder mehreren Gruppen vorliegt oder ob eine oder mehrere unabhängige Variablen einen signifikanten (nicht zufälligen) Einfluss auf eine abhängige Variable hat/haben. An dieser Stelle sei darauf verwiesen, dass **Forschungsprojekte i.d.R. immer inferenzstatistische Auswertungen erfordern**. Die Voraussetzung für die Wahl des richtigen Auswertungsverfahrens ist das Wissen über die Skalenniveaus der unabhängigen und abhängigen Variablen. Die Skalenniveaus wurden in Kapitel 5.1 erläutert und sind in der folgenden Tabelle noch einmal dargestellt:

Skalenniveau		Math. Eigenschaften	Beschreibung der Messwertcharakteristika	Beispiel	Lageparameter
Kategorial	Nominal	=/≠	Die Messwerte sind identisch oder nicht identisch.	Geschlecht	• Modus
	Ordinal	=/≠ ; </>	Die Messwerte sind größer, kleiner oder identisch.	Olympische Ränge	• Modus • Median
Metrisch	Intervall	=/≠ ; </> ; −; +	Die Distanz zwischen den Messwerten kann angegeben werden.	Temperatur	• Modus • Median • Arithm. Mittel
	Verhältnis	=/≠ ; </> ; +/− ; */÷	Die Distanz zwischen den Messwerten und deren Verhältnis kann angegeben werden.	Körpergröße	

Schauen Sie sich nun das **Video-Tutorial** zu den statistischen Analyseverfahren an:

Zu Beginn einer „statistischen Karriere“ fällt es vielen schwer, zwischen unabhängiger und abhängiger Variablen zu unterscheiden. Folgende Faustregel kann helfen: **die unabhängige Variable beeinflusst die abhängige Variable**, zum Beispiel: Wie stark beeinflussen die Werbeausgaben (unabhängige Variable) den Absatz eines Produktes (abhängige Variable)?

Die folgende Abbildung gibt einen Überblick über die einzelnen Verfahren und welches Verfahren bei welchen Skalenniveaus der unabhängigen und abhängigen Variablen angewandt wird:

Strukturprüfende Verfahren		**Unabhängige Variable**	
		nicht-metrisch	metrisch
Abhängige Variable	nicht-metrisch	Chi^2-Test	Diskriminanzanalyse
	metrisch	t-Test*** (bei 2 Gruppen) & Varianzanalyse (bei ≥ 3 Gruppen)	Regressionsanalyse (Dependenz) & Korrelationsanalyse (Interdependenz)

*** „Ergänzende“ Tests:
- Shapiro-Wilk-Test
  (prüft Stichproben auf Normalverteilung, die für den t-Test notwendig ist)
- Mann-Whitney-Test/Wilcoxon-Test
  (prüft mind. ordinal skalierte Daten auf Unterschiede ohne Berücksichtigung der Verteilungsform)

Strukturentdeckende Verfahren	• Hauptkomponentenanalyse (Faktoranalyse) • Clusteranalyse

Im Folgenden werden zur Verdeutlichung beispielhafte Forschungsfragen gegeben, die mit den einzelnen Analyse-Verfahren untersucht werden könnten:

- Chi^2-Test: Haben Frauen und Männer eine unterschiedliche Lieblingsfarbe (Unterscheidung in unabhängige und abhängige Variable ist aus statistischer Sicht nicht notwendig)?

- t-Test (zwei Kontrastgruppen): Haben Frauen und Männer (unabhängige Variable) ein unterschiedliches Einkommen (abhängige Variable)?
- Varianzanalyse (ANOVA) (mehr als zwei Kontrastgruppen): Haben die Bewohner der 16 Bundeländer (unabhängige Variable) ein unterschiedliches Einkommen (abhängige Variable)?
- Shapiro-Wilk-Test & Wilcoxon-Test:
    - wie t-Test, werden aber nur bei kleinen Stichproben eingesetzt (Faustregel: Stichprobenumfang < 50)
    - Shapiro-Wilk-Test prüft, ob die Stichprobe normalverteilt ist
    - Wilcoxon-Test prüft – wie der t-Test – Unterschiede zwischen zwei Gruppen
- Korrelationsanalyse: Gibt es einen (signifikanten / nicht zufälligen) Zusammenhang zwischen Alter und Einkommen (bei diesem Test gibt es keine abhängigen und unabhängigen Variablen)?
- Regressionsanalyse: Wie stark beeinflussen die Werbeausgaben (unabhängige Variable) den Absatz eines Produktes (abhängige Variable)?
- Hauptkomponentenanalyse (strukturentdeckendes Verfahren ohne abhängige und unabhängige Variablen): Können mehrere metrische Items (abgefragte Eigenschaften) zu Komponente („neuen Variablen“) zusammengefasst werden?

6.2 Chi-Quadrat-Test

Der Chi^2-Test prüft, ob nominale (kategoriale) Variablen (un-)abhängig voneinander sind – in anderen Worten: ob die gemessenen Werte nicht zufällig von den erwarten Werten (Werte, die bei „vollkommener“ Unabhängigkeit der Variablen gegeben sein müssten) abweichen. Folgende **Fragestellung** liegt ein Chi^2-Test zugrunde:

- Gibt es einen „nicht-zufälligen“ statistischen Zusammenhang zwischen zwei nominalskalierten Merkmalen? Sind zwei kategoriale Variablen unabhängig?

Eine konkrete Forschungsfrage könnte zum Beispiel lauten:

- Haben Frauen und Männer eine unterschiedliche Lieblingsfarbe (Unterscheidung in unabhängige und abhängige Variable ist aus statistischer Sicht nicht notwendig)?

Das **Vorgehen bei der Analyse** besteht aus zwei Schritten:

1) Ist der p-value <0.05 (<0.01, <0.001)? → Signifikanz?
2) Wie unterscheiden sich die Häufigkeiten?

Auch die **„statistischen" Analyseschritte („im Hintergrund")** bestehen aus zwei Schritten:

1) Berechnung der erwarteten Häufigkeiten (e), wenn kein Unterschied zwischen den Variablen bestehen würde (Spaltensumme * Zeilensumme / Merkmalsträger)
2) Bestimmung, ob die gemessenen Werte (h) signifikant von den erwarteten Werten abweichen

Schauen Sie sich in jedem Fall das **Video-Tutorial** zum Chi^2-Test an, da in diesem auch die Grundprinzipien der Inferenzstatistik ausführlicher erklärt werden, die auch allen weiteren Verfahren zugrunde liegen:

Aufgaben:

- Besteht ein signifikanter Unterschied zwischen Männern und Frauen bei der Präferenz für die SenderCluster?
- Unterscheidet sich das junge und alte Alterssegment bei der Affinität zu aktueller Musik? Und wenn ja, wie fällt dieser Unterschied aus?

Benötigte Befehle:

Chi^2 test

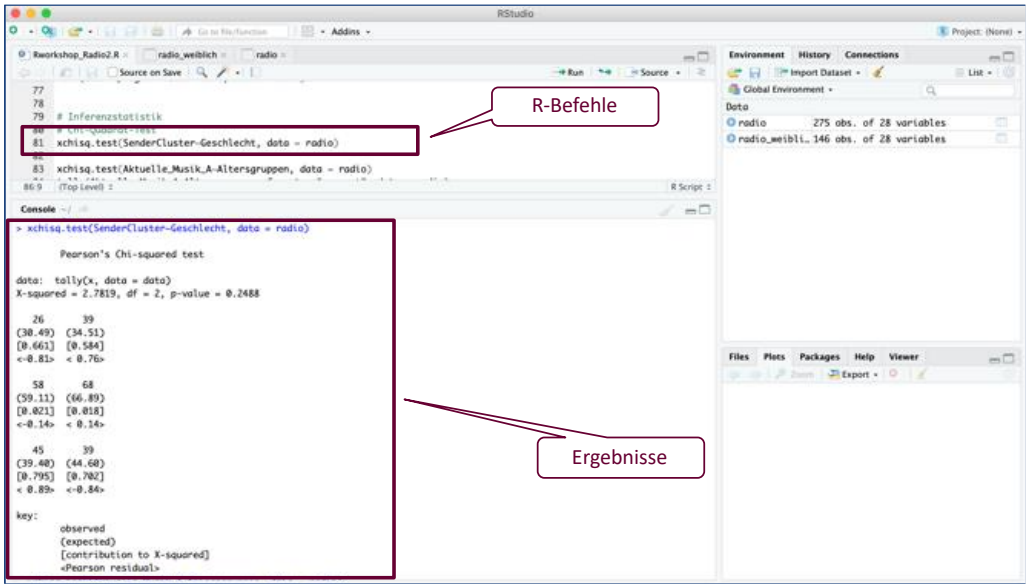
```
xchisq.test(variable1~variable2, data = dataset)
```

Relevante Variablen:

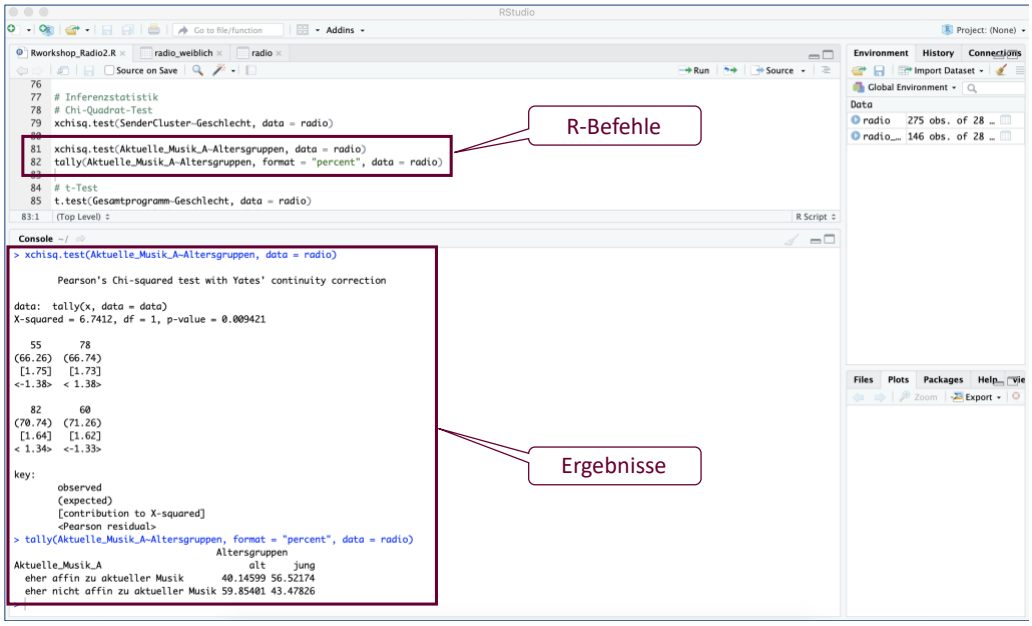
- SenderCluster
- Geschlecht
- Alter
- Altersgruppen
- Aktuelle_ Musik_A

Lösung in R:

R-Befehle und Ergebnisse zu Aufgabe 1:



R-Befehle und Ergebnisse zu Aufgabe 2:



6.3 t-Test

Der t-Test ist ein parametrisches Verfahren. Beim Zweistichproben-t-Test wird überprüft, ob sich eine metrische Variable zwischen zwei Stichproben signifikant (nicht zufällig) unterscheidet. Folgende **Fragestellungen** liegen einem t-Test zugrunde:

- Gibt es einen „nicht-zufälligen“ statistischen Unterschied zwischen zwei Stichproben in Bezug auf ein metrisches Merkmal?
- Gibt es einen „nicht-zufälligen“ Einfluss einer nominalen Variablen (2 Gruppen) auf eine metrische Variable?

Eine konkrete Forschungsfrage könnte zum Beispiel lauten:

- Haben Frauen und Männer (unabhängige Variable) ein unterschiedliches Einkommen (abhängige Variable)?

Der t-Test kann weiterhin **differenziert** werden in:

- t-Test bei unabhängigen Stichproben – Messungen bei zwei Gruppen
- t-Test bei abhängigen Stichproben – Messungen bei derselben Gruppe (z.B. vorher – nachher)

Das **Vorgehen bei der Analyse** besteht aus zwei Schritten (wie auch beim Chi^2-Test):

- Ist der p-value <0.05 (<0.01, <0.001)? → Signifikanz?
- Wie unterscheiden sich die Mittelwerte?

Im Folgenden finden Sie das **Video-Tutorial**, in dem der t-Test erklärt werden und wie man diesen in R anwendet:

Aufgaben:

- Beurteilen Männer und Frauen das Gesamtprogramm von Radiosendern (signifikant) unterschiedlich?
- Beurteilen Männer und Frauen die Musik von Radiosendern (signifikant) unterschiedlich?

Benötigte Befehle:

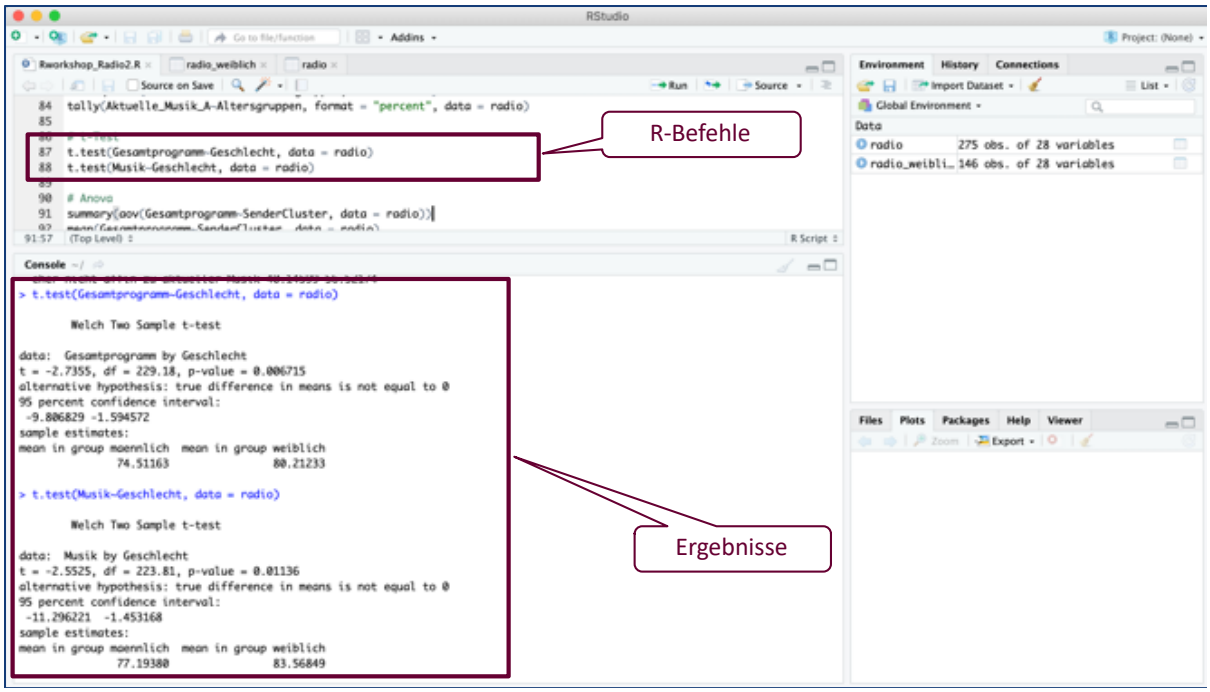
t-test

t.test(variable~grouping_variable, data = dataset)

Relevante Variablen:

- Musik
- Gesamtprogramm
- Geschlecht

Lösung in R:



6.4 ANOVA

Die Varianzanalyse (Analysis of Variance, ANOVA) ist ein parametrisches Verfahren und untersucht, ob sich die Mittelwerte mehrerer (mehr als zwei) unabhängiger Stichproben systematisch unterscheiden. Folgende **Fragestellung** liegt der ANOVA zugrunde:

- Gibt es einen „nicht-zufälligen“ statistischen Unterschied zwischen mehr als zwei Stichproben in Bezug auf ein metrisches Merkmal?

Eine konkrete Forschungsfrage könnte zum Beispiel lauten:

- Haben die Bewohner der 16 Bundeländer (unabhängige Variable) ein unterschiedliches Einkommen (abhängige Variable)?

Die ANOVA weist somit Ähnlichkeiten zum t-Test auf, lässt sich aber, wie folgt, vom **t-Test abgrenzen**:

- Während der t-Test metrische Merkmale von maximal zwei Stichproben auf signifikante Unterschiede prüft, werden bei der Varianzanalyse mehrere Mittelwerte verglichen.

Hier finden Sie das **Video-Tutorial** mit Aufgaben in R und Lösungen:

Aufgaben:

- Beurteilen die Hörer der verschiedenen SenderCluster das Gesamtprogramm von Radiosendern (signifikant) unterschiedlich?
- Beurteilen die Hörer der verschiedenen SenderCluster die Moderation von Radiosendern (signifikant) unterschiedlich?

Benötigte Befehle:

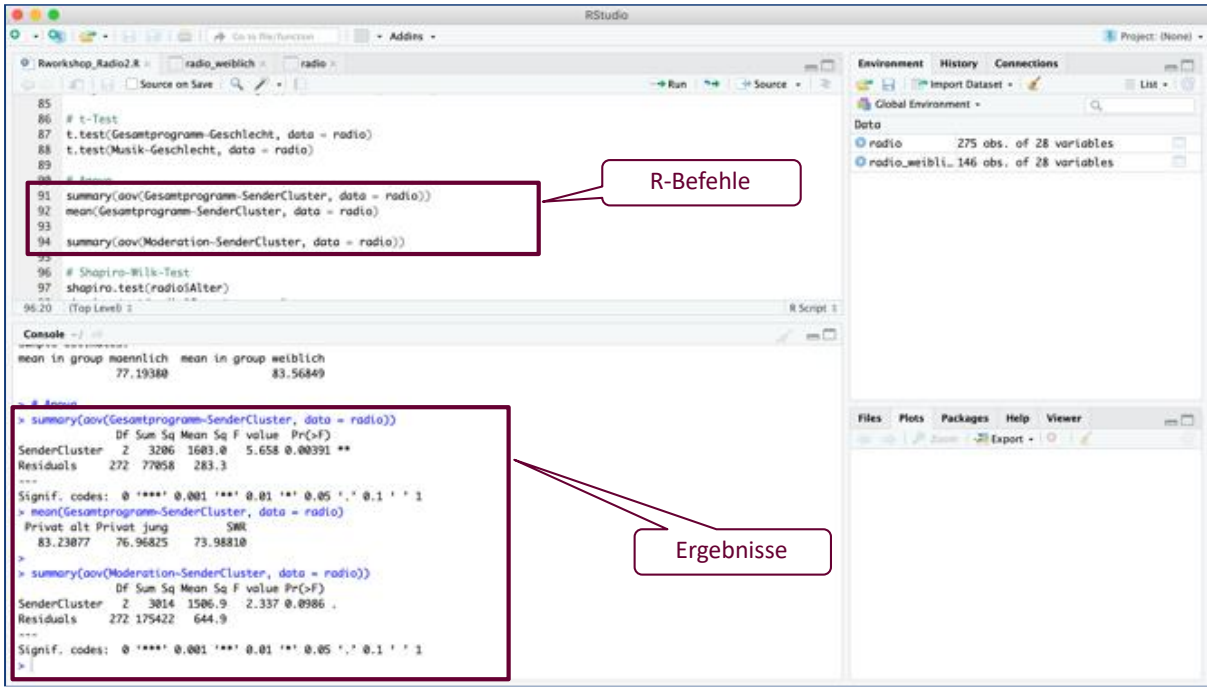
ANOVA

```
summary(aov(variable~grouping_variable, data = dataset))
```

Relevante Variablen:

- SenderCluster
- Moderation
- Gesamtprogramm

Lösung in R:



6.5 Shapiro-Wilk-Test

Verwendung:

Der Shapiro-Wilk-Test **vergleicht** die **Werte einer Stichprobe** mit **normalverteilten Werten**, um zu überprüfen, ob eine Normalverteilung angenommen werden kann. **Bei kleinen Stichproben (ca. n < 50) ist die Normalverteilung Voraussetzung für die Anwendung des t-Tests**. Folgende **Fragestellung** liegt dem Shapiro-Wilk-Test zugrunde:

- Sind die Werte einer mindestens intervallskalierten Variable normalverteilt?

Die **Ergebnisse des Shapiro-Wilk-Test** werden, wie folgt, **interpretiert**:

- Ist der Test **nicht signifikant** (p > .10), **unterscheidet sich die Verteilung** der vorliegenden Stichprobe **nicht von einer normalverteilten Stichprobe (→ Normalverteilung)**
- Ist der Test **signifikant** (p ≤ .10), **unterscheidet sich die Verteilung** der vorliegenden Stichprobe **von einer normalverteilten Stichprobe (→ keine Normalverteilung)**

Sollte bei einer kleinen Stichprobe der Shapiro-Wilk-Test ergeben, dass **keine Normalverteilung** vorliegt, ergibt sich die folgende **Konsequenz**:

- U.a. werden statt des t-Tests Rangverfahren wie Wilcoxon-/Mann-Whitney-Test angewendet (siehe nächstes Kapitel 6.6)
- Beachte: I.d.R. ist es sinnvoll, nicht nur den Shapiro-Wilk-Test durchzuführen, sondern auch einen Blick auf die (graphische) Verteilung zu werfen.

Schauen Sie sich nun das **Video-Tutorial** zum Shapiro-Wilk-Test an und lösen die darin gestellten Aufgaben mit R:

Aufgaben:

- Ist die Variable Alter normalverteilt?
- Ist die Variable Gesamtprogramm normalverteilt?

Benötigte Befehle:

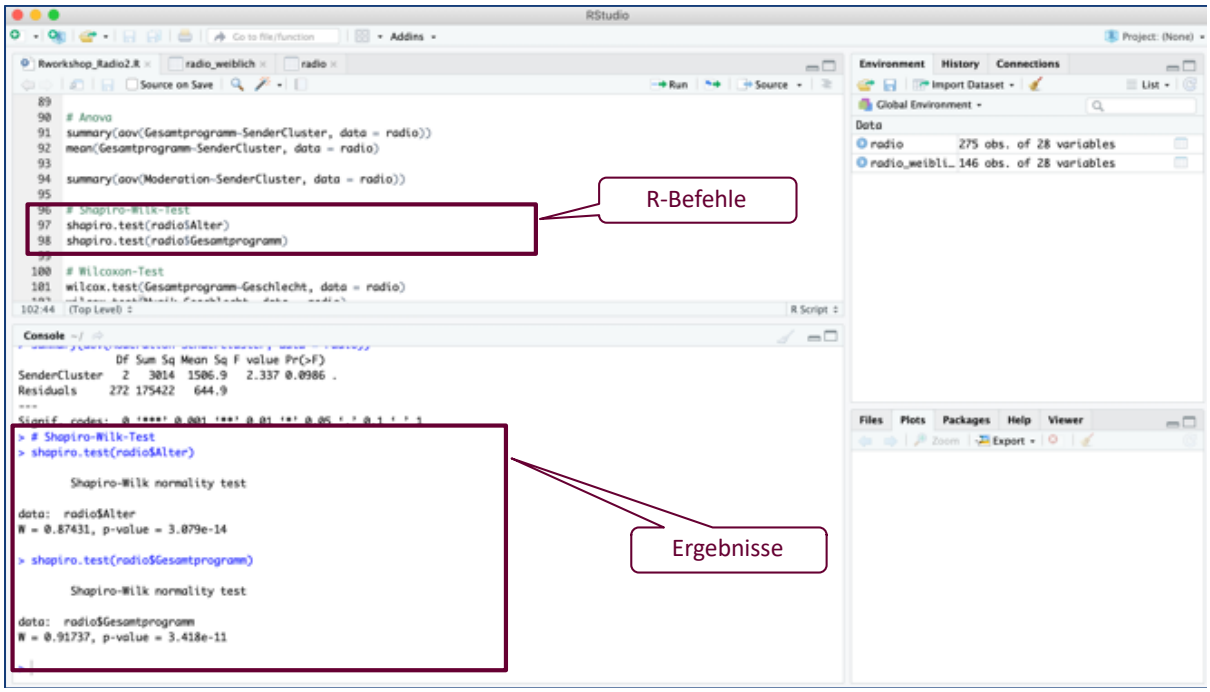
Shapiro-Wilk test

```
shapiro.test(dataset$variable)
```

Relevante Variablen:

- Gesamtprogramm
- Alter

Lösung in R:

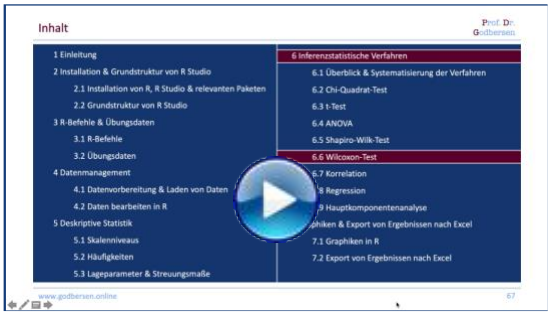


6.6 Wilcoxon-Test

Der Rangsummen-Test ist ein non-parametrisches Verfahren und vergleicht die *mittleren Rangzahlen* zweier Gruppen. (Es liegt keine Verteilungsannahme vor, u.a. Normalverteilung wurde auf Grund von Shapiro-Wilk-Test abgelehnt.) Folgende **Fragestellung** liegt dem Rangsummen-Test zugrunde:

- Gibt es einen „nicht-zufälligen“ statistischen Unterschied zwischen zwei Gruppen in Bezug auf ein mindestens ordinalskaliertes Merkmal?

Schauen Sie sich nun das **Video-Tutorial** zum Wilcoxon-Test an und lösen die darin gestellten Aufgaben mit R:



Aufgaben:

- Liegt nach dem Wilcoxon-Test ein signifikanter Unterschied zwischen Männern und Frauen bei der Beurteilung des Gesamtprogramms vor?
- Liegt nach dem Wilcoxon-Test ein signifikanter Unterschied zwischen Männern und Frauen bei der Beurteilung der Musik vor?

Benötigte Befehle:

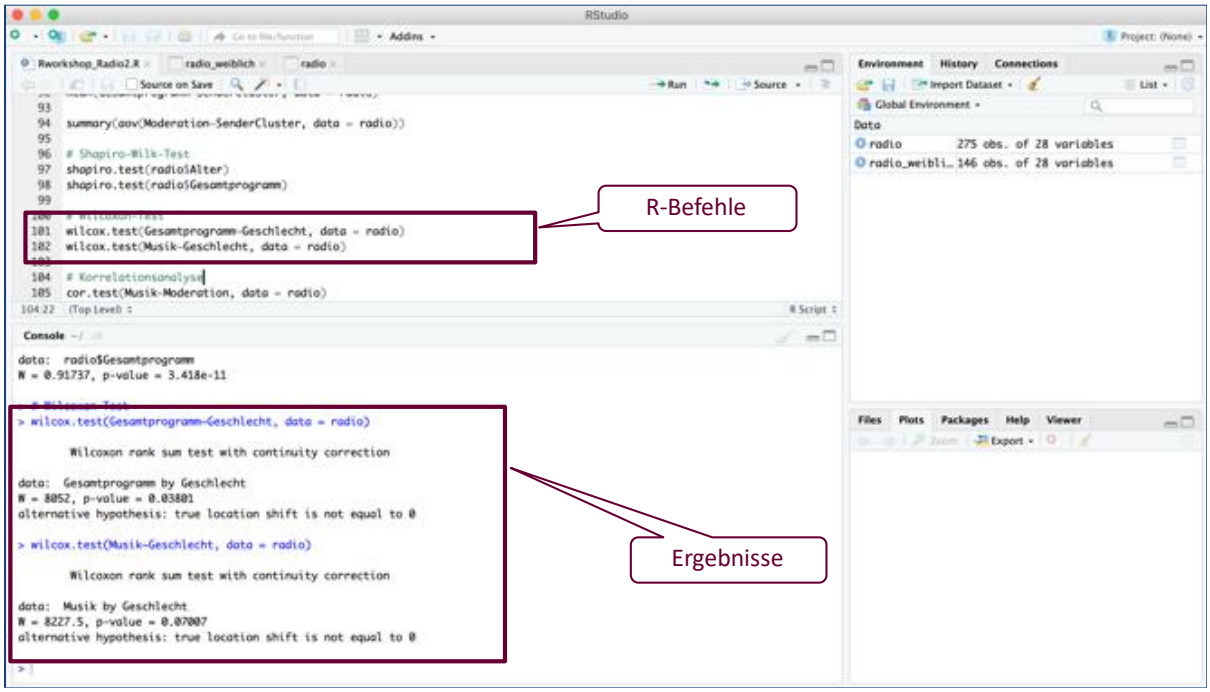
Mann-Whitney & Wicoxon test

```
wilcox.test(variable~grouping_variable, data = dataset)
(→Mann-Whitney test)
```

Relevante Variablen:

- Musik
- Gesamtprogramm
- Geschlecht

Lösung in R:



6.7 Korrelation

Die Korrelationsanalyse untersucht den (linearen) Zusammenhang zwischen zwei oder mehr metrischen Variablen und beantwortet die folgende **Fragestellung**:

- Wie stark ist der lineare Zusammenhang zwischen den Variablen – und in welcher Richtung besteht er? → Wird der Wert von Variable A erhöht / gesenkt oder bleibt er gleich, wenn sich der Wert von Variable B ändert?

Eine konkrete Forschungsfrage könnte zum Beispiel lauten:

- Gibt es einen (signifikanten / nicht zufälligen) Zusammenhang zwischen Alter und Einkommen (bei diesem Test gibt es keine abhängigen und unabhängigen Variablen)?

Das **Vorgehen bei der Analyse** besteht aus zwei Schritten (vgl. auch Kap. 6.2 Chi2-Test und 6.3 t-Test):

1) Ist der p-value <0.05 (<0.01, <0.001)? → Signifikanz?
2) Wie hoch ist der Korrelationskoeffizient?

Eine **mögliche Interpretation des Korrelationskoeffizienten** finden Sie im Folgenden:

- $0.2 < |r| \leq 0.5$: geringe Korrelation
- $0.5 < |r| \leq 0.7$: mittlere Korrelation
- $0.7 < |r| \leq 1.0$: hohe Korrelation

Dabei ist das Vorzeichen des Korrelationskoeffizienten zu beachten:

- $0 < r \leq +1.0$: gleichgerichteter Zusammenhang
- $-1.0 \geq r < 0$: entgegengesetzter Zusammenhang

Im folgenden **Video** wird die Korrelationsanalyse genauer erklärt. Schauen Sie deshalb auf jeden Fall dieses Video und lösen Sie die darin gestellten Aufgaben:

Aufgaben:

- Gibt es einen nicht-zufälligen Zusammenhang zwischen der Beurteilung der Musik und der Moderation?
- Gibt es einen nicht-zufälligen Zusammenhang zwischen der Beurteilung von Comedy und Gewinnspielen?

Benötigte Befehle:

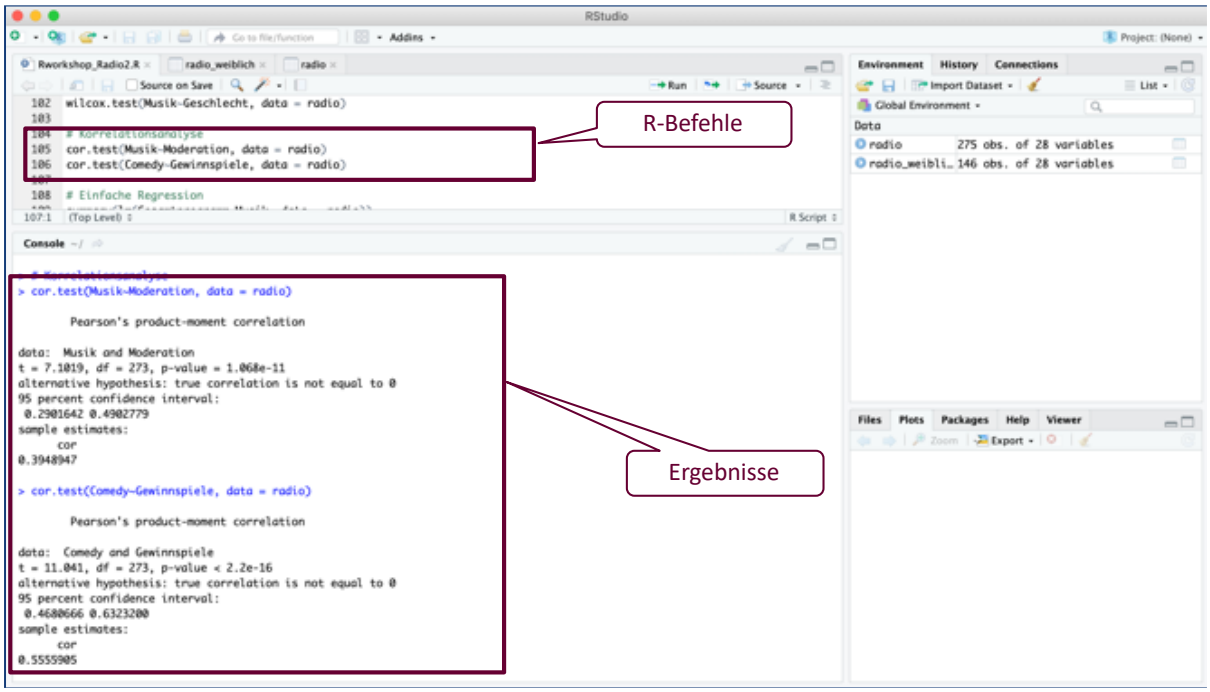
Correlation

```
cor.test(variable1~variable2, data = dataset)
```

Relevante Variablen:

- Musik
- Moderation
- Comedy
- Gewinnspiele

Lösung in R:



6.8 Regression

Die lineare Regression untersucht den linearen Einfluss einer metrischen Variablen (Einfachregression) oder mehrerer metrischer Variablen (Multiple Regression) auf eine abhängige metrische Variable. Dabei wird folgende **Fragestellung** beantwortet:

- Inwieweit kann der Wert einer abhängigen Variablen durch die Werte einer (Einfachregression) oder mehrere (Multiple Regression) unabhängigen Variablen erklärt werden?

Eine konkrete Forschungsfrage könnte zum Beispiel lauten:

- Wie stark beeinflussen die Werbeausgaben (unabhängige Variable) den Absatz eines Produktes (abhängige Variable)?

Das **Ergebnis der Analyse** ist die **Regressionsgerade** (Ergebnis der Analyse): **y = a + b * x**

- y: Abhängige Variable („Wirkung“)
- x: Unabhängige Variable („Ursache“)
- a: y-Achsenabschnitt
- b: Regressionskoeffizient („um wieviel steigt y, wenn x sich um eine Einheit erhöht“)

Statistische Analyse („im Hintergrund“): Im Rahmen der Regressionsanalyse wird die Gerade ermittelt, bei der die quadrierten Abweichungen der gemessen Werte von der Geraden am geringsten ist (Methode der kleinsten Quadrate)

Das **Vorgehen bei der Analyse** besteht aus drei Schritten:

1) Sind die p-values <0.05 (<0.01, <0.001)?
   - Liegt ein signifikanter Einfluss von x_i auf y vor (p-values)?
1) Wie hoch sind die Regressionskoeffizienten (b)?
   - Wie stark ist der Einfluss von xi auf y?
2) Welchen Wert hat das Bestimmtheitsmaß (R^2)?
   - Wie hoch ist der Erklärungsgehalt des Modells?
   - R^2 = Anteil der Varianz der abhängigen Variablen, den die unabhängigen Variablen erklären

Schauen Sie sich nun das **Video-Tutorial** zur Regressionsanalyse mit einer genaueren Erklärung des Verfahrens sowie den Übungsaufgaben und deren Lösung in R an:

Aufgaben:

- Welchen Einfluss hat die Musik auf das Gesamtprogramm?
- Welchen Einfluss hat die Comedy auf das Gesamtprogramm?
- Welchen Einfluss haben die fünf Programmelemente auf das Gesamtprogramm?

Benötigte Befehle:

Linear regression

```
summary(lm(dependent_variable~independent_variable, data = dataset))
```

Multiple regression

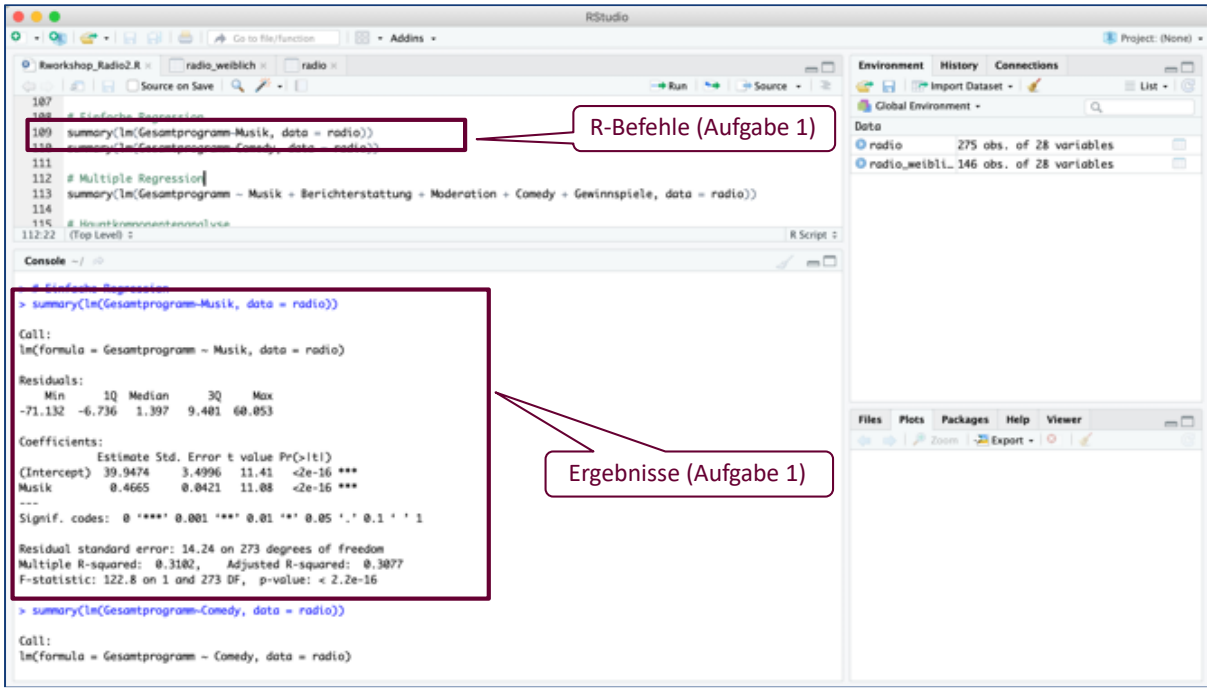
```
summary(lm(dependent_variable~ind_variable1 + ind_variable2, data = dataset))
```

Relevante Variablen:

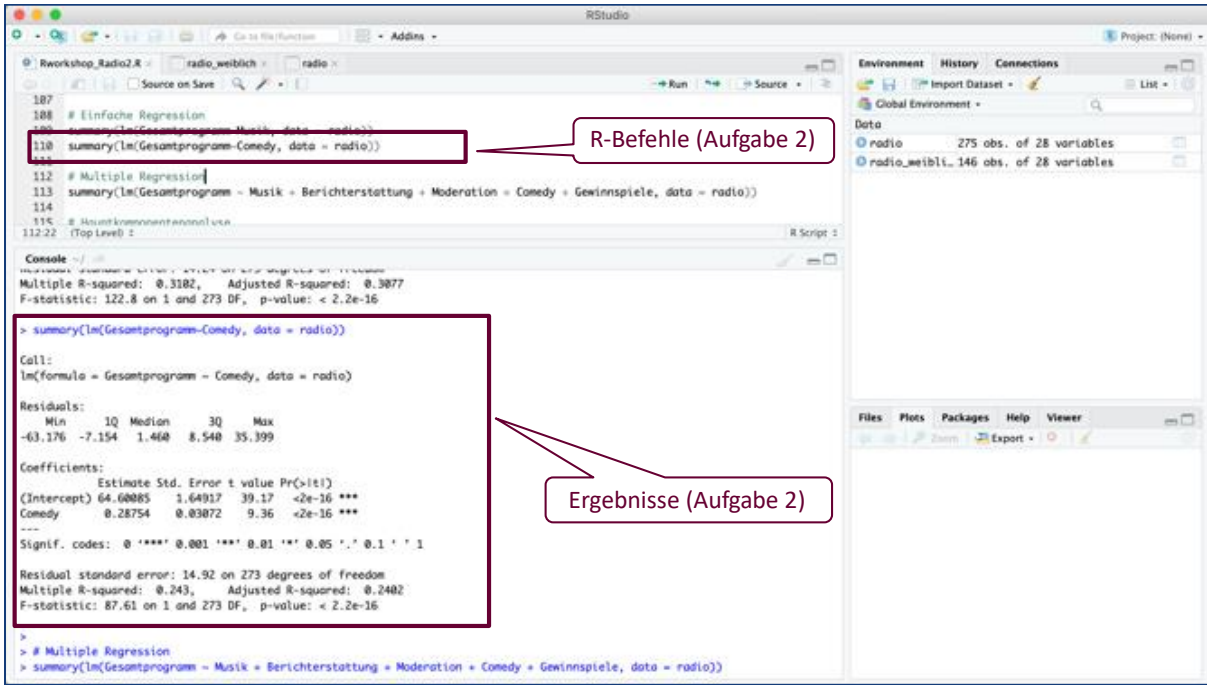
- Berichterstattung
- Musik
- Moderation
- Comedy
- Gewinnspiele
- Gesamtprogramm

Lösung in R:

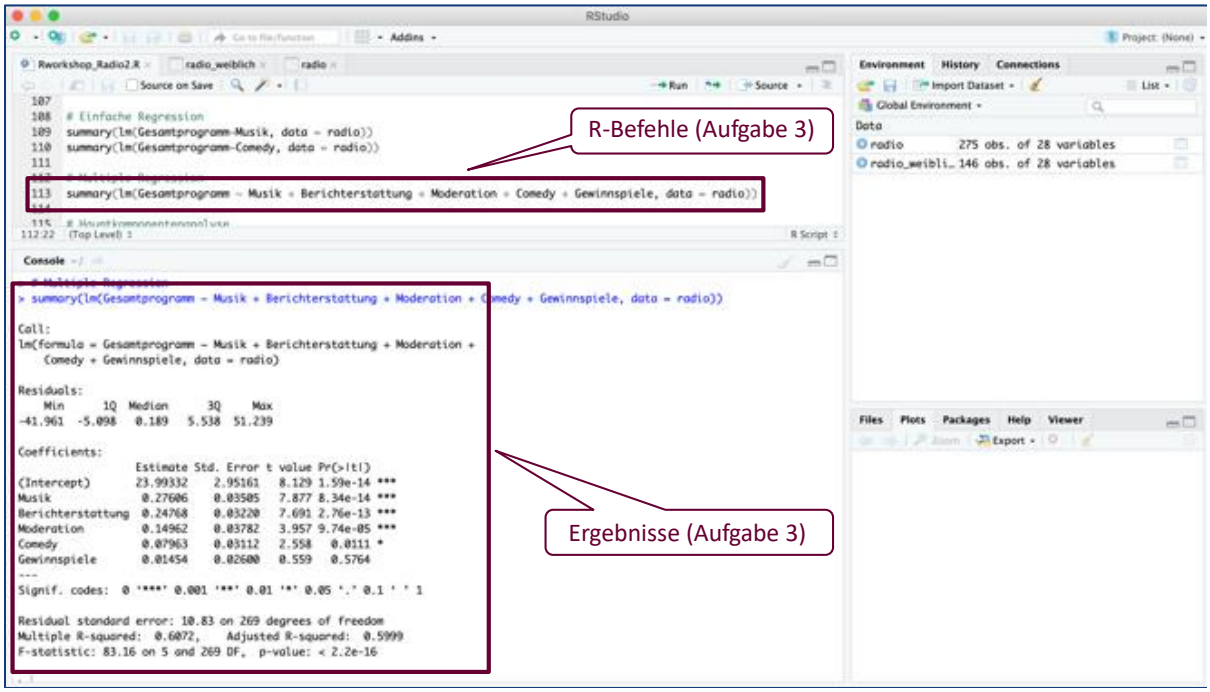
R-Befehle und Ergebnisse zu Aufgabe 1:



R-Befehle und Ergebnisse zu Aufgabe 2:



R-Befehle und Ergebnisse zu Aufgabe 3:



6.9 Hauptkomponentenanalyse

Die Hauptkomponentenanalyse wird verwendet, um Fragen/Variablen/Indikatoren zu wenigen Dimensionen zusammenzufassen, und beantwortet die folgende **Fragestellung**:

- Können multidimensionale metrische Daten auf wenige wichtige Hauptkomponenten verdichtet werden?

Die **Grundidee** der Hauptkomponentenanalyse ist:

- Die Items einer Komponente korrelieren hoch miteinander.
- Die Items unterschiedlicher Komponenten korrelieren nicht hoch miteinander.

Die **Umsetzung der Hauptkomponentenanalyse** in R erfolgt ein einem Vorbereitungsschritte und dann fünf Analyseschritten:

0) Vorbereiten der Datenbasis (neuer Datensatz)
1) Prüfung, ob eine Hauptkomponentenanalyse durchgeführt werden kann/soll mittels KMO (Kaiser-Meyer-Olkin-Test: misst den Anteil der Varianz zwischen Variablen, der auf gemeinsame Varianz zurückzuführen ist)

2) Bestimmung der Anzahl der Komponenten anhand der Eigenwerte
   - Auswahl: So viele Komponenten wie Eigenwerte > 1
3) Bestimmung der Faktorladungen (Korrelation eines Items mit einer Komponente) & Zuordnung der Items zu Komponenten
   - Voraussetzung: Ladung > 0.5
   - Entscheidung: Zuordnung des Items zu der Komponente, auf der es am höchsten lädt
4) Interpretation der Komponenten
5) Testen der einzelnen Komponenten mittels Cronbachs $\boldsymbol{\alpha}$ (Reliabilität; durchschnittliche Korrelation aller Split-half-tests)

Im Folgenden **Video-Tutorial** wird die Hauptkomponentenanalyse im Detail und an einem Beispiel erklärt; ferner wird in dem Video auch die Übungsaufgabe gestellt und die dazugehörige Lösung in R erläutert:

Aufgaben:

- Zu welchen Komponenten lassen sich die Programmeigenschaften zusammenfassen? („Wie können aus den Programmeigenschaften neue übergeordnete Variablen gebildet werden?“)
- Wie ist die Reliabilität der neuen Komponenten zu beurteilen?

Benötigte Befehle:

Principal Component analysis (package: psych)

```
dataset_new <- cbind(dataset$variable1, dataset$variable2…)
colnames(dataset_new) <- c("Item_1", "Item_2")
KMO(dataset_new)
pcaX <- princomp(dataset_new, scores = TRUE, cor = TRUE)
summary(pcaX)                    plot(pcaX)
principal(dataset_new, nfactors = x, rotate = “varimax”)
```

Cronbach’s Alpha (package: psy)

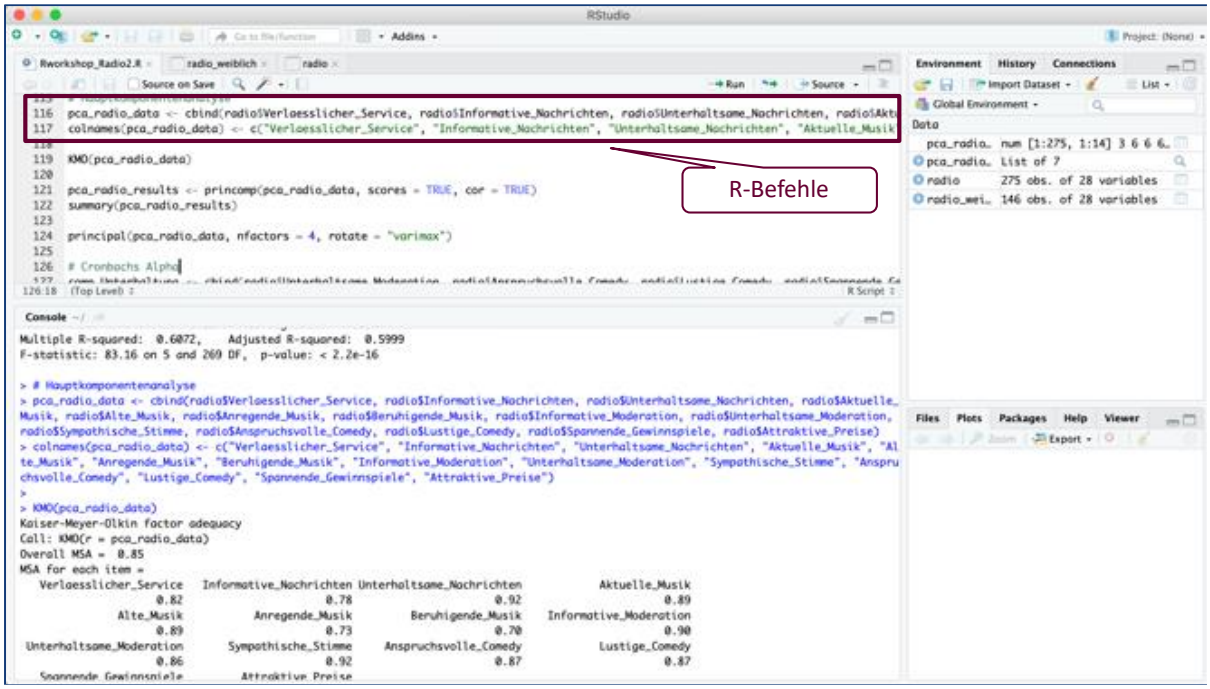
```
factorX <- cbind(dataset$variable1, dataset$variable2…)
cronbach(factorX)
```

Relevante Variablen:

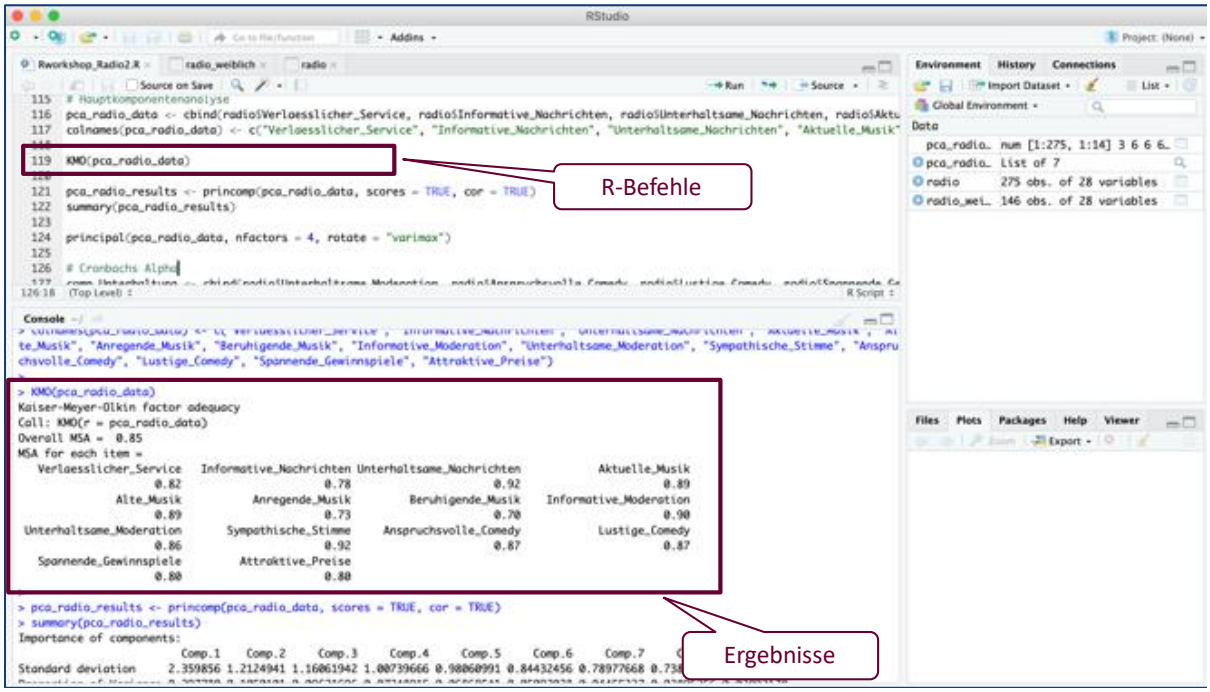
- Verlaesslicher_Service
- Informative_Nachrichten
- Unterhaltsame_Nachrichten
- Aktuelle_Musik
- Alte_Musik
- Anregende_Musik
- Beruhigende_Musik
- Informative_Moderation
- Unterhaltsame_Moderation
- Sympathische_Stimme
- Anspruchsvolle_Comedy
- Lustige_Comedy
- Spannende_Gewinnspiele
- Attraktive_Preise

Lösung in R entlang der fünf Schritte der Hauptkomponentenanalyse:

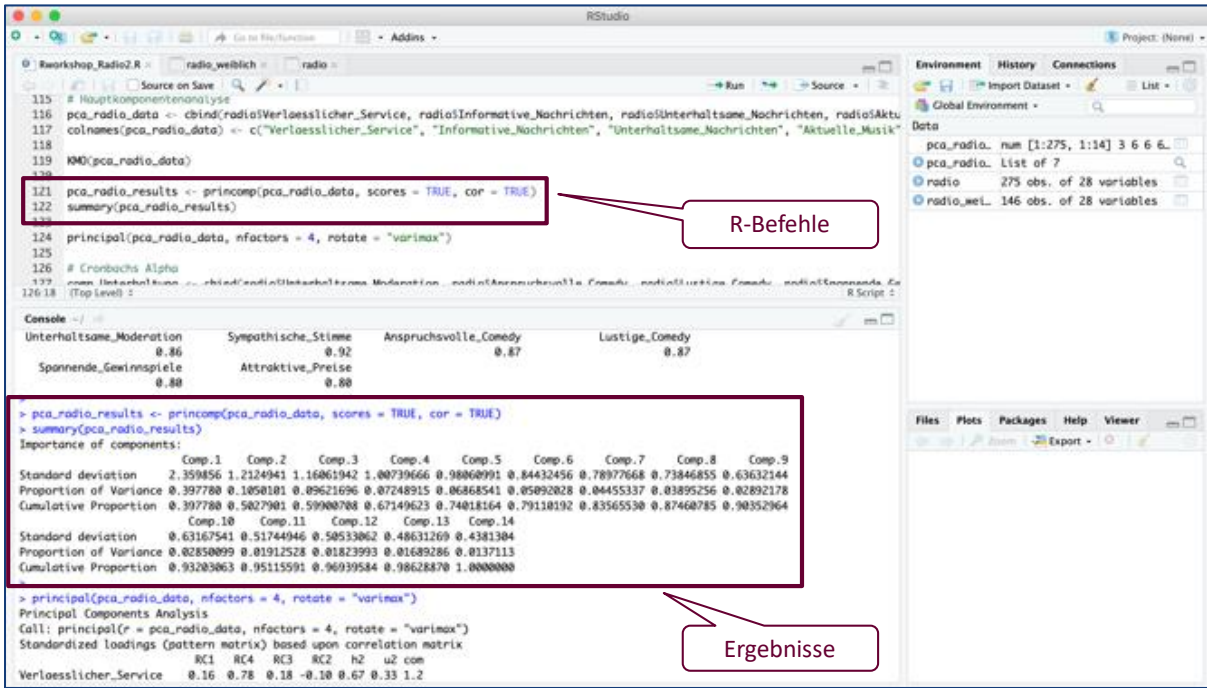
0) Vorbereiten der Datenbasis (neuer Datensatz)



1) Prüfung mittels KMO

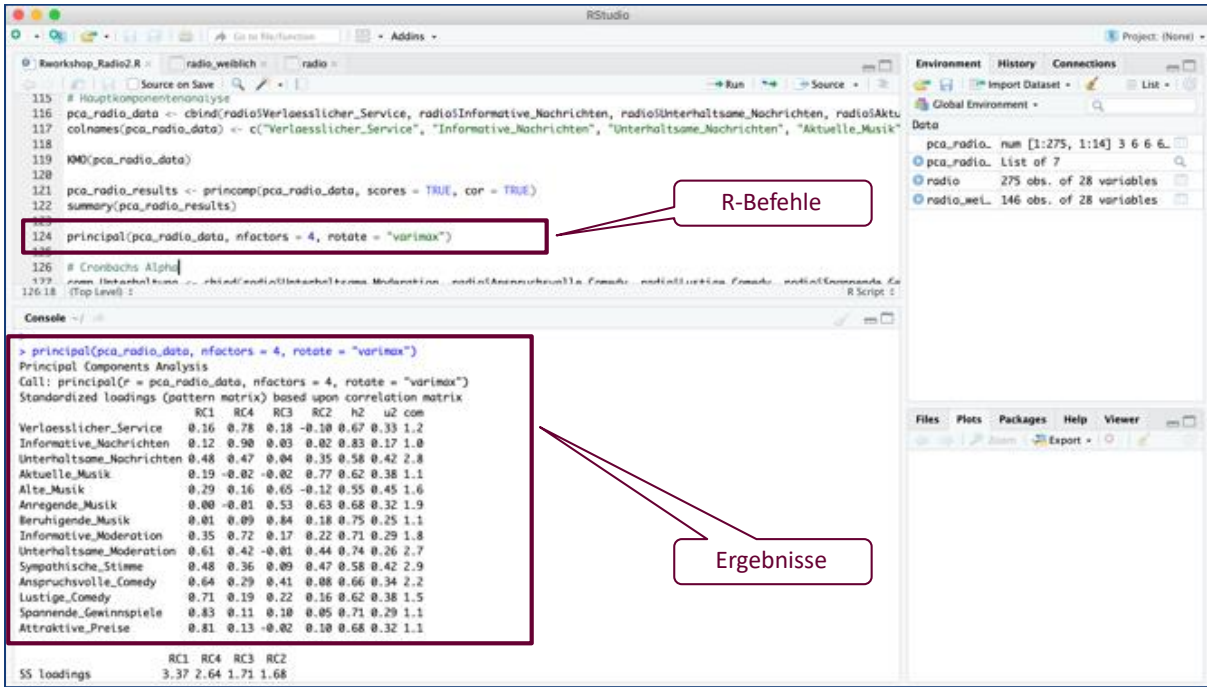


2) Bestimmung der Anzahl der Komponenten anhand der Eigenwerte

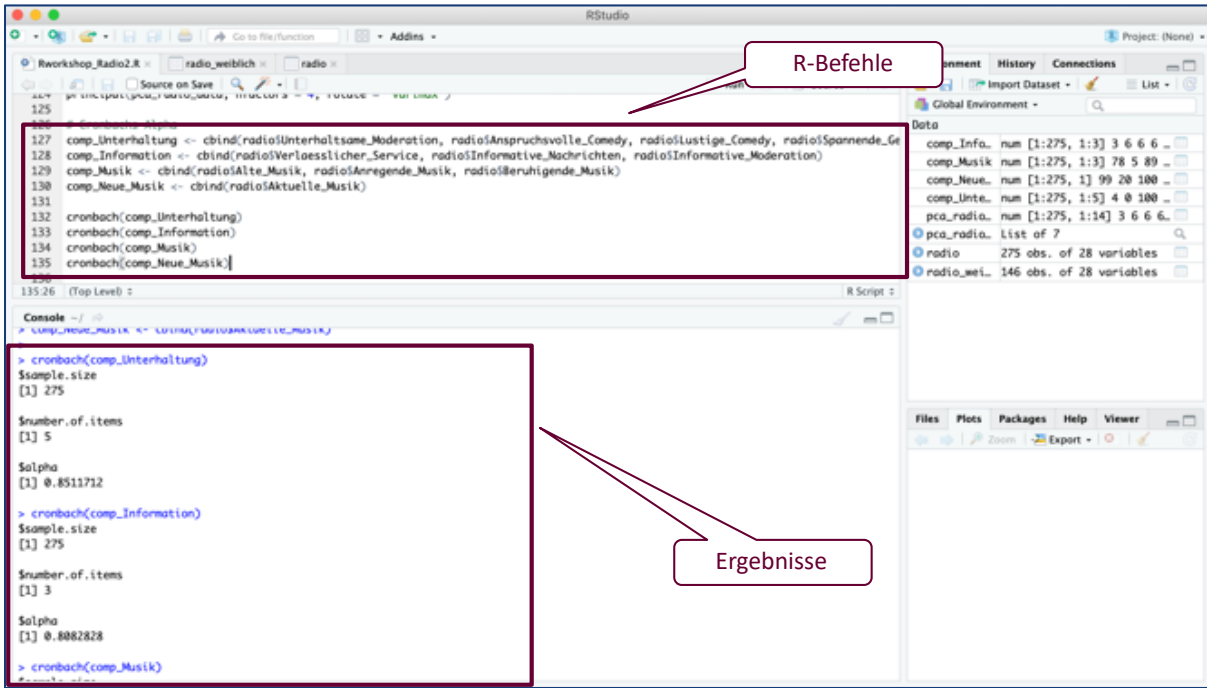


3) Bestimmung der Faktorladungen und

4) Interpretation der Komponenten



5) Testen der einzelnen Komponenten mittels Cronbachs α



7 Graphiken & Export von Ergebnissen nach Excel

Bei der Ergebnisdokumentation in einem Forschungsartikel, einer Abschlussarbeit oder einem Projektbericht ist es i.d.R. sinnvoll, Graphiken zu verwenden, um dem Leser eine leichtere Nachvollziehbarkeit zu ermöglichen. In den folgenden beiden Unterkapiteln lernen Sie, wie Sie Graphiken in R erstellen können und Ergebnisse nach Excel exportieren können, um dort Graphiken zu erstellen.

7.1 Graphiken in R

Im folgenden **Video** wird erläutert, wie Sie Ergebnisgraphiken in R erstellen:

Aufgaben:

- Erstellen Sie ein Säulendiagramm, das die relativen Häufigkeiten (%) der Variable Geschlecht widergibt.
- Erstellen Sie ein Säulendiagramm, dass die relativen Häufigkeiten (%) der Variable SenderCluster in Abhängigkeit vom Geschlecht widergibt.

Benötigte Befehle:

Bar graph – frequencies with 1 variable

```
table <- tally(~variable, format = "proportion",data = dataset)
frame <- as.data.frame(table)
View(frame)
ggplot(frame, aes(variable, Freq)) + geom_bar(position = "dodge", stat = "identity", fill = "#143C78") +
scale_y_continuous(limits = c(min, max)) + geom_text(aes(label = sprintf("%0.2f", frame$Freq)), position =
position_dodge(width = 0.9), vjust = -0.25, colour = "black", size = 2.7)
```

Bar graph – frequencies with 2 variables

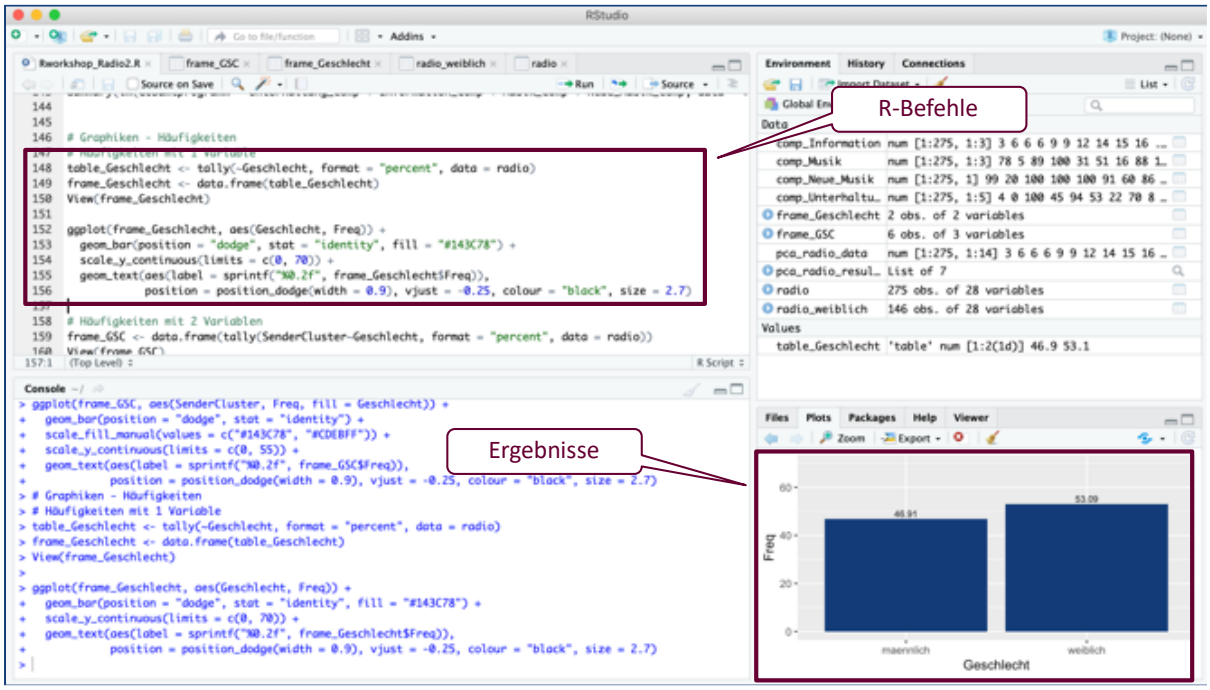
```
table <- tally(variable_1~variable_2, format= "proportion",data = dataset)
frame <- as.data.frame(table)
View(frame)
ggplot(frame, aes(variable_1, Freq, fill = variable_2)) + geom_bar(position = "dodge", stat = "identity") +
scale_fill_manual(values = c("#143C78", "#CDEBFF")) + scale_y_continuous(limits = c(min, max)) + geom_text(aes(label =
sprintf("%0.2f", frame$Freq)), position = position_dodge(width = 0.9), vjust = -0.25, colour = "black", size = 2.7)
```

Relevante Variablen:

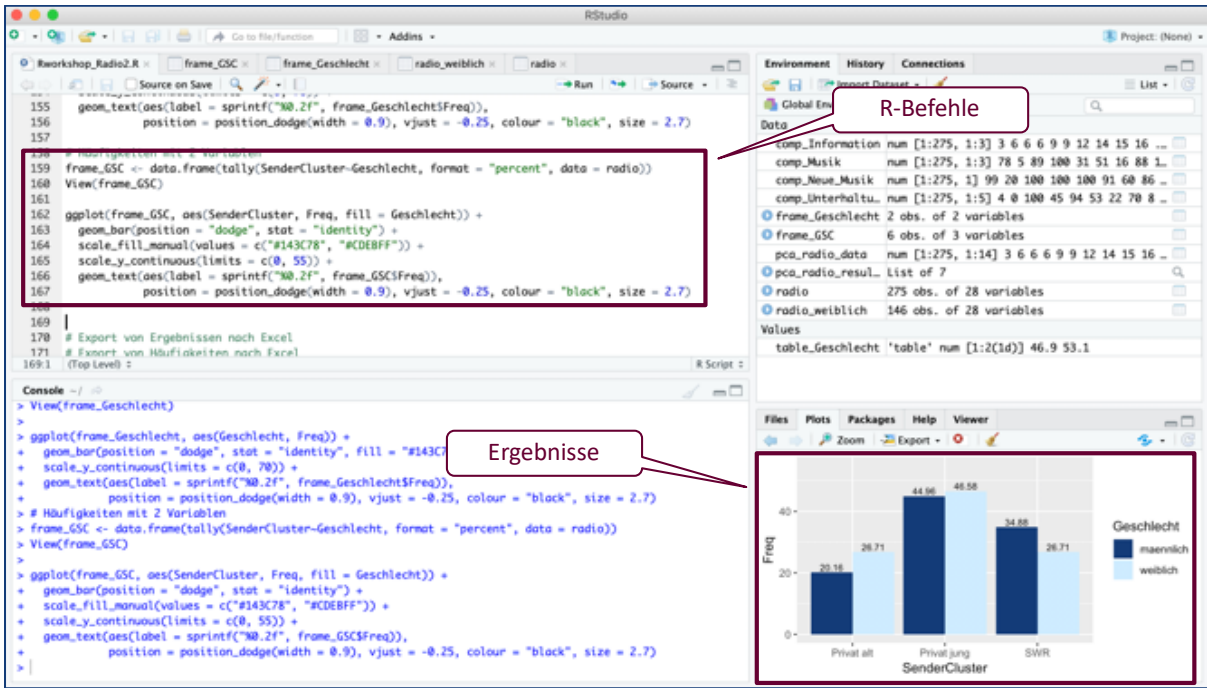
- SenderCluster
- Geschlecht

Lösung in R:

R-Befehle und Ergebnisse zu Aufgaben 1:

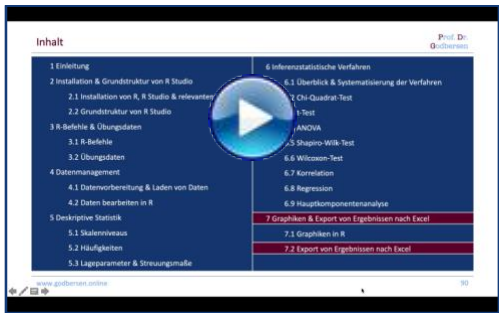


R-Befehle und Ergebnisse zu Aufgabe 2:



7.2 Export von Ergebnissen nach Excel

Im folgenden **Video** wird erläutert, wie Sie Analyseergebnisse aus R nach Excel exportieren können, um dort Ergebnisgraphiken zu erstellen:



Aufgaben:

- Exportieren Sie die relativen Häufigkeiten (%) der Variable SenderCluster in Abhängigkeit vom Geschlecht nach Excel.
- Exportieren Sie die Mittelwerte der Bewertungen der Programmelemente nach Excel.

Benötigte Befehle:

Export R results to Excel (package: openxlsx)

(1.1) Preparing frequencies

```
frame <- data.frame(tally(variable_1~variable_1, data = dataset))
```

(1.2) Preparing mean, median etc.

```
frame <- data.frame(new_variable_name_1 =
c(mean(~variable_1, data = dataset), mean(~variable_2, data =
dataset)), new_variable_name_2 = c(mean(~variable_3, data =
dataset), mean(~varaible_4, data = dataset)))
```

(2) Export to Excel

```
write.xlsx(frame, "Desktop/R_operations/name.xlsx", asTable =
FALSE)
```

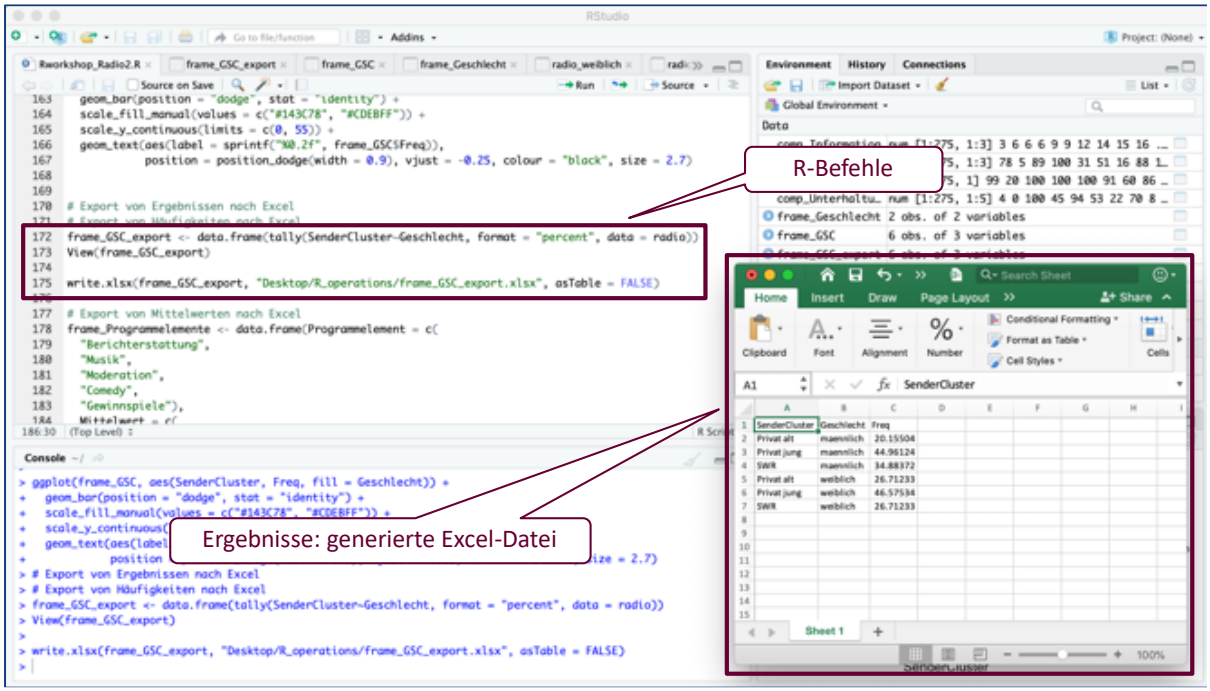
→ Please note: The name of the R frame & the directory / name of the xlsx-file have to be adjusted

Relevante Variablen:

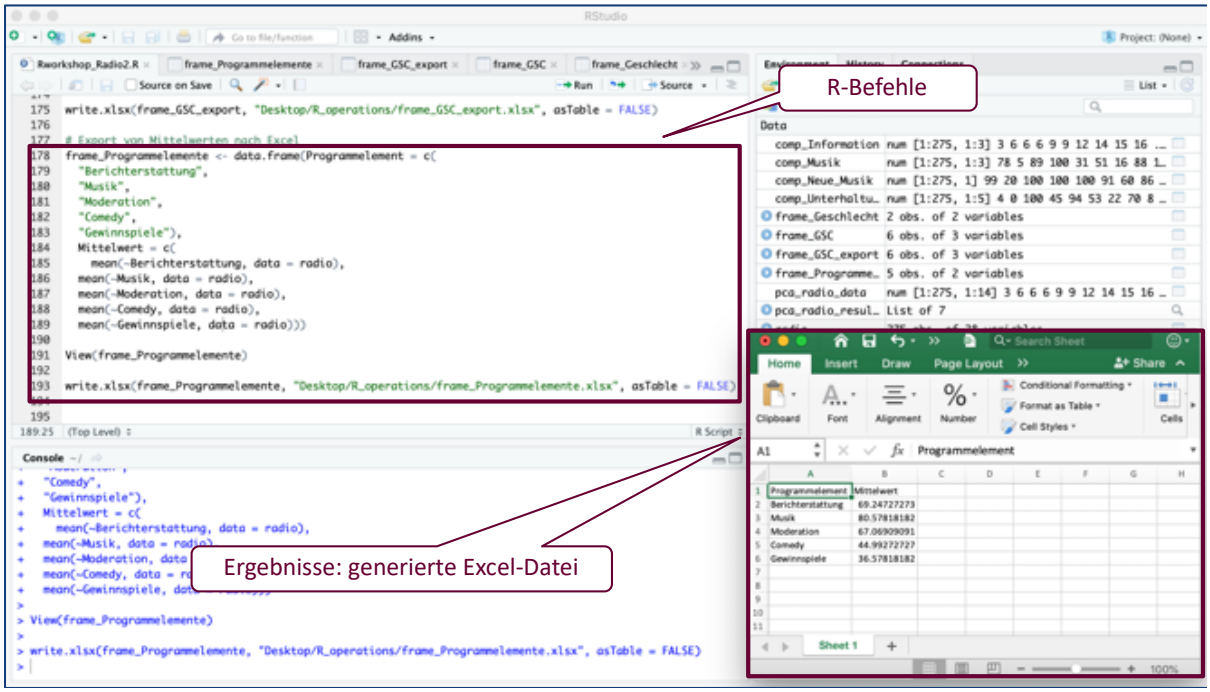
- SenderCluster
- Berichterstattung
- Musik
- Moderation
- Comedy
- Gewinnspiele
- Gesamtprogramm
- Geschlecht

Lösung in R:

R-Befehle und Ergebnisse zu Aufgabe 1:



R-Befehle und Ergebnisse zu Aufgabe 2:



Abschluss des Kurses und Nachwort

Sie sind am Ende des Kurses Angewandte Statistik & R angelangt. Wenn Sie die einzelnen Kapitel Schritt für Schritt durchgegangen sind und alle Übungsaufgaben lösen konnten, sollten Sie nun in der Lage sein, ein empirisches Forschungsprojekt erfolgreich auszuwerten.

Falls Sie Ihre neu gewonnenen Kenntnisse und Fähigkeiten festigen wollen, können Sie alle Übungsaufgaben noch einmal in R bearbeiten. Legen Sie auch hier den Fokus auf die beiden Fragen: Wann wende ich welches Verfahren an und wie interpretiere ich die Ergebnisse?

Der Autor würde sich über Ihr Feedback zu diesem Kurs freuen. Senden Sie dazu eine Email an: hendrik.godbersen@godbersen.online.

Ich wünsche Ihnen viel Erfolg bei Ihrem nächsten empirischen Projekt.

Ihr Prof. Dr. Hendrik Godbersen