Das DHV*Lab* Analysecenter

Ein Tool zur datengestützten Forschung in den Geisteswissenschaften

Stefanie Schneider

Digitale Forschungswerkzeuge breiten sich auf mannigfaltige Weise in den Geisteswissenschaften aus; seien es nun webbasierte geografische Informationssysteme oder digitale historisch-kritische Editionen.[1] Nichtsdestotrotz werden teils ausgefeilte und nützliche *Tools* kaum eingesetzt.[2] Eine Studie, in der Fred Gibbs und Trevor Owens überwiegend US-amerikanische Historiker befragten, welche Bedürfnisse *Tools* zu erfüllen haben, nannte dafür zwei wesentliche Gründe: Erstens gehen viele *Tools* nicht auf den traditionellen geisteswissenschaftlichen Anwender ein und stellen keine einfachen und selbsterklärenden Oberflächen bereit. Zweitens mangelt es ihnen häufig an Anleitungen mit praxisnahen Beispielen und an einer Dokumentation, die den methodologischen Wert des *Tools*, und seine Limitierungen, mit nichttechnischem Vokabular beschreibt.[3]

Das im Folgenden präsentierte *Analysecenter*[4] nimmt sich dieser Kritikpunkte an. Auf der einen Seite stellt es eine prädefinierte, flexible

1 *Digital Research Tools* (*DiRT*) bietet eine Übersicht verschiedenster Werkzeuge in den unterschiedlichsten Kategorien, beispielsweise „Analyze relationships between pieces of data" und „Transcribe audio, video or manuscripts". Siehe http://dirtdirectory.org/ (09.08.2017).

2 „Only about six percent of humanist scholars go beyond general purpose information technology and use digital resources and more complex digital tools in their scholarship." Siehe Summit on Digital Tools in the Humanities: A Report on the Summit on Digital Tools, Charlottesville 2005, S. 4.

3 Siehe Gibbs, Fred und Owens, Trevor: Building Better Digital Humanities Tools: Toward Broader Audiences and User-Centered Designs, in: Digital Humanities Quarterly 6.2, 2012, http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html (09.08.2017). Gibbs und Owens schreiben weiterhin (Abschnitt 7): „Tool interfaces must help more traditional historians feel more comfortable with new ways of visualizing, analyzing, and thinking about sources and about data."

4 http://dhvlab.gwi.uni-muenchen.de/app/analysis/ (09.08.2017).

URN: http://nbn-resolving.de/urn:nbn:de:bvb:19-epub-42417-6

Werkzeugpalette zur Verfügung, um einen Gegenstand, die *Small* bis *Large* oder *Biggish Data*,[5] mit einer intuitiven grafischen Benutzeroberfläche zu untersuchen und die Ergebnisse dieser Untersuchung, Grafiken und Tabellen, komfortabel zu exportieren. Auf der anderen Seite ist es als Modul in das *Digital Humanities Virtual Laboratory*[6] integriert und nutzt die dort bestehende Infrastruktur – darunter Manuale, die in die Statistik und Informatik mit konkreten, aus den Geisteswissenschaften stammenden Anwendungsfällen einführen –, auch in der Lehre.[7] Das *Analysecenter* hat zwei Ziele: erstens eine zuvor definierte Fragestellung zu beantworten, zweitens eine in diesem Zuge neue Fragestellung zu gewinnen; das Bekannte zu untermauern und das Unbekannte zu explorieren – mithilfe der Empirie, sowohl auf Mikro- als auch auf Makroebene.[8] Die Programmiersprache *R*[9], das auf *R* aufsetzende Webapplikationspaket *shiny*[10] sowie die Erweiterungen *shinydashboard*[11] und *shinyjs*[12] bilden dafür das technische *Framework*. Das sogenannte *tidyverse*[13] unterstützt ferner einen kohärenten *Workflow* im *Backend*; es organisiert, strukturiert, transformiert und manipuliert den Gegenstand. Als *Data* des momentanen Prototyps fungieren rund sieben Millionen *Taggings* der kunsthistori-

5 Da *Big Data* nochmals andere Herausforderungen mit sich bringt, die das *Analysecenter* nicht berücksichtigt, wird der Begriff bewusst unterlassen.

6 Eine „digitale Lehr- und Forschungsinfrastruktur für die Datenanalyse in den Kunst-, Geschichts- und Sprachwissenschaften" der Ludwig-Maximilians-Universität München; siehe die weiteren Beiträge dieses Bandes und ferner http://dhvlab.gwi.uni-muenchen.de/ (09.08.2017).

7 Das *Analysecenter* war unter anderem Gegenstand der Seminare „Big Data in der Kunstgeschichte" im Wintersemester 2016/2017 und „Automatische und manuelle Bildadressierung in der Kunstgeschichte" im Sommersemester 2017 an der Ludwig-Maximilians-Universität München (beide Hubertus Kohle und Stefanie Schneider).

8 Dazu auch Weaver, Warren: Science and Complexity, in: American Scientist 36, 1948, S. 536–544.

9 R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2017, https://www.r-project.org/ (09.08.2017).

10 Chang, Winston et al.: shiny: Web Application Framework for R, 1.0.3, 2017, https://cran.r-project.org/package=shiny (09.08.2017).

11 Chang, Winston und Borges Ribeiro, Barbara: shinydashboard: Create Dashboards With Shiny, 0.6.1, 2017, https://cran.r-project.org/package=shinydashboard (09.08.2017).

12 Attali, Dean: shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds, 0.9.1, 2017, https://cran.r-project.org/package=shinyjs (09.08.2017).

13 Wickham, Hadley: tidyverse: Easily Install and Load Tidyverse, 1.1.1, 2017, https://cran.r-project.org/package=tidyverse (09.08.2017).

schen Spieleplattform *ARTigo*[14], die Spielern digitale Reproduktionen von Kunstwerken präsentiert, um diese zu annotieren.[15] Die *Taggings* eines konkret interessierenden Werks werden in den nächsten Absätzen näher betrachtet, um einige ausgewählte Funktionen des *Analysecenters* vorzustellen.

Wenden wir uns exemplarisch Tizians „Venus mit Cupido, Orgelspieler und Hündchen" (Abb. 1) und den diesem Werk ähnlichen Werken zu. Der erste Reiter des *Analysecenters*, Übersicht, zeigt links alle Künstler und die Anzahl ihrer in *ARTigo* annotierten Kunstwerke (Abb. 2). Tizian findet sich auf einer hinteren Seite, und seine „Venus mit Cupido, Orgelspieler und Hündchen" schließlich im mittleren *Container*, wenn links die Zeile angewählt wird, die das *Label* „Titian" trägt.[16] Ein Klick auf „Venus mit Cupido, Orgelspieler und Hündchen" gibt rechts in einem weiteren *Container* die *Tags* und ihre absolute Häufigkeit, die sogenannte *Suchwortdichte*, frei: darunter zunächst die elementaren Schlagwörter „Frau", „Hund", „Mann" und „Orgel", die jeweils viermal annotiert wurden, doch danach ebenso die aus kunsthistorischer Perspektive spezifischeren *Tags* „Putte", zweimal, sowie „Venus" und „Renaissance", jeweils einmal (Abb. 3). Die quantitative Analyse beginnt im *Browsing*, in einer deskriptiven *Tour d'Horizon* durch den Gegenstand, der ihm anhaftenden Variablen und deren Ausprägungen.[17]

Rücken wir näher an die Fragestellung. Der dritte Reiter des *Analysecenters*, Ähnlichkeitsanalysen, fächert zwei Unterreiter, *1-Alle-Vergleich* und *Alle-Alle-Vergleich*, auf. Zugrunde liegt ihnen das im *Information Retrieval* populäre Modell, einen Text zu *tokenisieren*, ihn also in seine Bestandteile, Wörter – oder allgemeiner Terme –, zu zerlegen, und diese derart zu sortieren und zu indexieren, dass ein Vokabular

14 http://www.artigo.org/ (09.08.2017).

15 Für eine detaillierte Beschreibung siehe Kohle, Hubertus: Kunstgeschichte Goes Social Media, in: Aviso: Zeitschrift für Wissenschaft und Kunst in Bayern 3, 2011, S. 38–43.

16 Die alternative Schreibweise, Titian statt Tizian, ist zu beachten und in der Such- und Filtermaske über einen *regulären Ausdruck*, beispielsweise „Ti(z|t)ian", abzufangen.

17 Dazu Chang, Shan-Ju und Rice, Ronald E.: Browsing: A Multidimensional Framework, in: Annual Review of Information Science and Technology 28, 1993, S. 231–276.

entsteht, und mit ihm eine mathematische Struktur, die sich für die quantitative Analyse eignet.[18] Tizians „Venus" wandelt sich auf diese Weise in einen Vektor, der – repräsentiere er lediglich die zuvor spezifizierten sieben Tags – durch

$$r = (w_1, w_2, w_3, w_4, w_5, w_6, w_7) = (4, 4, 4, 4, 2, 1, 1)$$

definiert wird, wobei *r* die sogenannte *Ressource* bezeichne, Tizians „Venus mit Cupido, Orgelspieler und Hündchen", und w_1 bis w_7 die *Suchwortdichte* der über Indizes referenzierten sieben *Tags* wiedergebe, „Frau", „Hund", „Mann", „Orgel", „Putte", „Venus" und „Renaissance", aus denen sich die *Ressource* im mathematischen Sinne bildet. Dieses Modell ist aus zwei Gründen praktikabel: Zum einen spannt ein Vektor einen Raum auf,[19] der weitere Vektoren enthalten kann; ein Beispiel ist Redons „Geburt der Venus". Zum anderen liegen nicht nur ähnlich konstellierte *Ressourcen*[20] in diesem euklidischen Raum nahe beieinander, sondern es ist auch ihre Nähe zu messen, und zwar über den Kosinus des zwischen ihnen liegenden Winkels. Hier setzt die uns interessierende Ähnlichkeit an.

Selektieren wir nun Tizians „Venus mit Cupido, Orgelspieler und Hündchen" im Reiter *1-Alle-Vergleich*.[21] Die Resultate der ersten Seite (Abb. 4) illustrieren: Zuvorderst erscheint eine im *Museo del Prado* ausgestellte Wiederholung der „Venus" – noch mit Orgel und Orgel-

18 Salton, Gerard, Wong, Anita und Yang, Chung-Shu: A Vector Space Model for Automatic Indexing, in: Communications of the ACM 18.11, 1993, S. 613.

19 Die mathematisch präzisere, doch weniger anschauliche Formulierung lautet: Ein Vektor bildet eine Basis eines Raums.

20 Ähnlich konstelliert meint in diesem Beispiel, aus vielen gleichen *Tags* zusammengesetzt, die ähnlich häufig annotiert wurden.

21 Unter *Lokale Gewichtung* wurde das Feld *Suchwortdichte*, unter *Globale Gewichtung* das Feld *Keine Gewichtung* und unter *Ähnlichkeitsmaß* das Feld *Kosinus-Ähnlichkeit* markiert. Weiterhin wurde die Anzahl der *Tags*, die beide Kunstwerke, das prädefinierte feste und das mit jeder Iteration wechselnde, gemein haben müssen, auf zehn reduziert, um die Menge der möglichen Ergebnisse zu erhöhen. Eine Einführung in die unterschiedlichen Maße, einen Term, auch in Abhängigkeit anderer Terme, zu gewichten, liefern Salton, Gerard und Buckley, Christopher: Term Weighting Approaches in Automatic Text Retrieval, in: Information Processing and Management 24.5, 1988, S. 513–523.

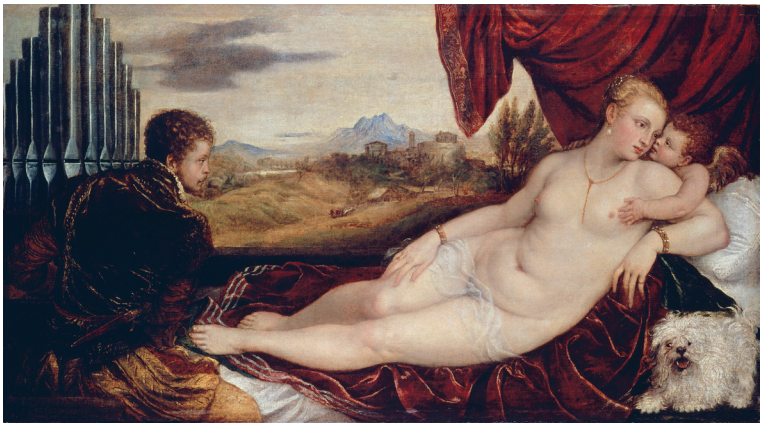

Abbildung 1: Tizian, „Venus mit Cupido, Orgelspieler und Hündchen“, um 1550, *Gemäldegalerie Berlin*. Gemeinfrei.

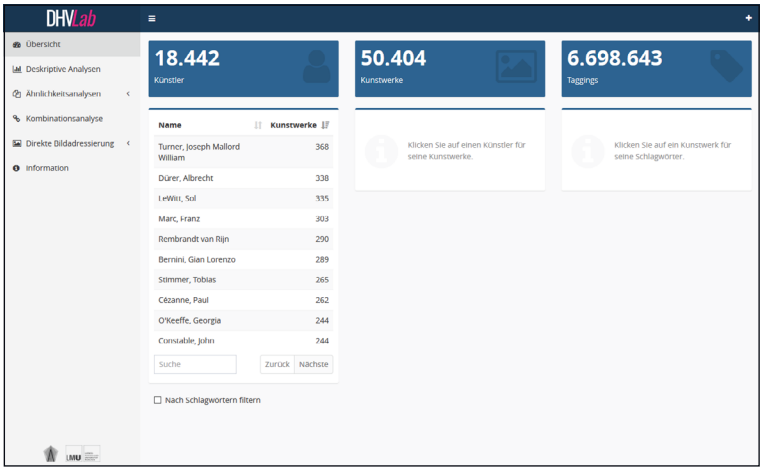

Abbildung 2: Reiter *Übersicht* des *Analysecenters*. Der *Container* links zeigt alle Künstler und die Anzahl der Kunstwerke des Künstlers, die in *ARTigo* annotiert wurden.

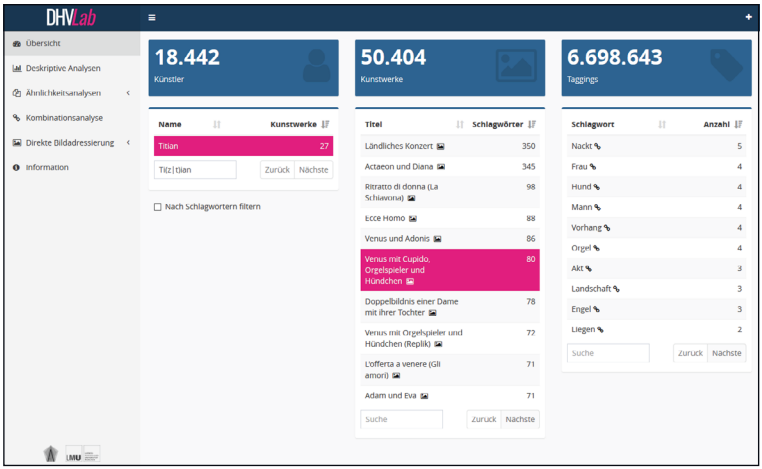

Abbildung 3: Reiter *Übersicht* des *Analysecenters*. Im linken *Container* wurde nun Tizian selektiert, im mittleren seine „Venus mit Cupido, Orgelspieler und Hündchen“. Der *Container* rechts zeigt alle für dieses Werk bislang hinterlegten *Tags*.

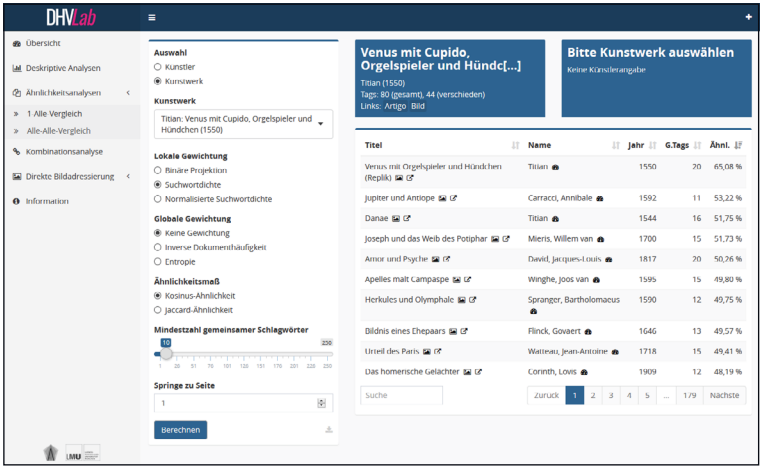

Abbildung 4: Reiter *1-Alle-Vergleich* des *Analysecenters*. Selektiert wurde Tizians „Venus mit Cupido, Orgelspieler und Hündchen“.

Abbildung 5: Tizian, „Venus mit Orgelspieler und Hündchen“, um 1550, *Museo del Prado*. Gemeinfrei. Vergleiche dazu auch Abb. 1.

Abbildung 6: Jacques Louis David, „Amor und Psyche“, um 1817, *Cleveland Museum of Art* (links), Lovis Corinth, „Das homerische Gelächter“, um 1909, *Neue Pinakothek München* (rechts). Beide gemeinfrei.

spieler, doch ohne Cupido und mit anderem Schoßhündchen (Abb. 5) –, die zweifellos auch unter traditionellen hermeneutischen Kriterien als zu der in der *Gemäldegalerie Berlin* ausgestellten „Venus“ ähnlich deklariert werden würde. Bereits an Position fünf und zehn folgen allerdings Davids „Amor und Psyche“ und Corinths „Das homerische Gelächter“, die allein ob ihres Titels, und, mehr noch, ihrer Datierung – „Amor und Psyche“ entstand um 1817, „Das homerische Gelächter“ um 1909 –, aus dem Raster des Naheliegenden fallen. Ein Blick auf die als sogenannte *Ausreißer* deklarierten Werke legitimiert ihr Auf-

treten teilweise. Beide zeigen ein nacktes Paar, Mann und Frau, auf einer Schlafstätte; die Venus nur zeigen sie beide nicht: Bei David ist es die schlafende Psyche, bei Corinth die in flagranti ertappte Aphrodite (Abb. 6). Weder Psyche noch Aphrodite werden jedoch, wenn es nach den Spielern von *ARTigo* geht, im bedeutsamen Maße erkannt. Nur ein *Tagging* weist „Psyche" direkt aus, während sich „Aphrodite" überhaupt nicht unter den hinterlegten *Taggings* findet.

Der durch die quantitative Analyse suggerierte Objektivismus ist kontextueller Natur.[22] Obgleich *gegeben*, interagiert der Gegenstand, Tizians „Venus mit Cupido, Orgelspieler und Hündchen", mit einem Subjekt, das ihn erst als *gegeben* konstituiert. Da es sich bei *ARTigo* um ein Spiel handelt, ist hier der Spieler das Subjekt, und der Spieler ordnet sich den Regeln des Spiels unter, um es möglichst erfolgreich zu absolvieren: Die Punkte, die er für die Eingabe eines gültigen, weil in einer anderen als der aktuellen Spielrunde annotierten *Tags* erhält,[23] beeinflussen die *Tags*, die er tätigt. Hierbei entsteht ein *Bias*. Wissen wir jedoch um ihn, deuten wir Davids „Amor und Psyche" und Corinths „Das homerische Gelächter" als Produkte des sie generierenden Prozesses und als Produkte des Algorithmus, der sie transformiert. Der hier den Gegenstand konstituierende Prozess leitet den Spieler an, *Tags* „trivialerer" Natur zu annotieren; schließlich birgt „Liege" eine höhere Wahrscheinlichkeit mit Punkten belohnt zu werden als beispielsweise „Chaiselongue". Der hier den Gegenstand transformierende Algorithmus wiederum sieht die einzelnen *Tags* als unabhängig zueinander. Eine Chaiselongue ist im semantischen Sinne aber eine Liege, und als solche sind sich auch die *Tags* „Chaiselongue" und „Liege" ähnlicher als beispielsweise die *Tags* „Chaiselongue" und „Horizont"; zumindest aus der Perspektive eines Menschen. Der in diesem Beispiel verwendete Algorithmus weiß darum nicht: Für ihn sind sich alle *Tags* in gleicher

22 Lisa Gitelman und Virginia Jackson bemerken ebenso: „The point is not how to judge whether objectivity is possible [...] but how to describe objectivity in the first place." (Gitelman, Lisa und Jackson, Virginia: Introduction, in: Gitelman, Lisa (Hrsg.): „Raw Data" is an Oxymoron, Cambridge 2013, S. 4.)

23 Das Prinzip des *Matching* dämmt Missbrauch ein. Ein einmal in Edvard Munchs Porträt des August Strindberg annotiertes „Micky Maus" etwa wird selten ein zweites Mal annotiert werden.

Weise unähnlich.[24] Ihn gilt es ebenso zu reflektieren wie den Gegenstand, dem er sich nähert.[25]

Deskription, Exploration und Reflexion rahmen den multiperspektivischen *Workflow* des *Analysecenters*. Dieser interaktive und iterative Prozess des Entdeckens ist essenziell, beide Ziele – eine zuvor definierte Fragestellung zu beantworten und eine in diesem Zuge neue Fragestellung zu gewinnen – zu realisieren, und ist damit gleichwohl in die universitäre Lehre zu integrieren. Ein *Tool* wie das hier anhand einer exemplarischen Fragestellung umrissene *Analysecenter* steht letztlich niemals entkoppelt: Die Algorithmen, die es implementiert, sind zwar nicht mehr eigenhändig zu programmieren. Auf theoretischer Ebene zu verstehen sind sie aber dennoch, um die aus der Arbeit mit dem *Tool* resultierenden Ergebnisse entsprechend, auch auf qualitativer Ebene, interpretieren und mögliche Limitierungen einschätzen zu können.

Literatur

Attali, Dean: shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds, 0.9.1, 2017, https://cran.r-project.org/package=shinyjs (09.08.2017).

Chang, Shan-Ju und Rice, Ronald E.: Browsing: A Multidimensional Framework, in: Annual Review of Information Science and Tech nology 28, 1993, S. 231–276.

Chang, Winston, Cheng, Joe, Allaire, Joseph J., Xie, Yihui und McPherson, Jonathan: shiny: Web Application Framework for R, 1.0.3, 2017, https://cran.r-project.org/package=shiny (09.08.2017).

Chang, Winston und Borges Ribeiro, Barbara: shinydashboard: Create Dashboards With Shiny, 0.6.1, 2017, https://cran.r-project.org/package=shinydashboard (09.08.2017).

24 An dieser Stelle sei angemerkt, dass es durchaus Algorithmen gibt, die semantische Relationen zwischen Wörtern adäquat modellieren können; beispielsweise *Latent Semantic Indexing*.

25 Siehe Stalder, Felix: Algorithmen, die wir brauchen, 2017, https://netzpolitik.org/2017/algorithmen-die-wir-brauchen/ (09.08.2017).

Gibbs, Fred und Owens, Trevor: Building Better Digital Humanities Tools: Toward Broader Audiences and User-Centered Designs, in: Digital Humanities Quarterly 6.2, 2012, http://www.digitalhumanities.org/dhq/vol/6/2/000136/000136.html (09.08.2017).

Gitelman, Lisa und Jackson, Virginia: Introduction, in: Gitelman, Lisa (Hrsg.): „Raw Data" is an Oxymoron, Cambridge 2013, S. 1–14.

Kohle, Hubertus: Kunstgeschichte Goes Social Media, in: Aviso: Zeitschrift für Wissenschaft und Kunst in Bayern 3, 2011, S. 38–43.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2017, https://www.r-project.org/ (09.08.2017).

Salton, Gerard und Buckley, Christopher: Term Weighting Approaches in Automatic Text Retrieval, in: Information Processing and Management 24.5, 1988, S. 513–523.

Salton, Gerard, Wong, Anita und Yang, Chung-Shu: A Vector Space Model for Automatic Indexing, in: Communications of the ACM 18.11, 1993, S. 613–620.

Stalder, Felix: Algorithmen, die wir brauchen, 2017, https://netzpolitik.org/2017/algorithmen-die-wir-brauchen/ (09.08.2017).

Summit on Digital Tools in the Humanities: A Report on the Summit on Digital Tools, Charlottesville 2005.

Weaver, Warren: Science and Complexity, in: American Scientist 36, 1948, S. 536–544.

Wickham, Hadley: tidyverse: Easily Install and Load Tidyverse, 1.1.1, 2017, https://cran.r-project.org/package=tidyverse (09.08.2017).