**Einführung in R & Shiny Apps:**

**Cancer App**



https://blog.efpsa.org/2019/04/24/7-easy-steps-to-building-your-own-shiny-app-from-scratch/

**Eine Gruppenarbeit von**

**Kussi Katsha Benjamin Michel Devrim Evelik**

**DSM 5/ 3132414 DSM 8/ 3127536 DSM 8/ 3127529**

**Einleitung**

Der Lungenkrebs ist eine menschenabhängige und auch eine sehr häufig auftretende Krebsart. Daher haben wir den Lungenkrebs auf unserem Projekt bezogen.

Dieser ist ein bösartiger Tumor, der von Bronchialschleimhautzellen oder Lungengewebezellen ausgeht. Ärzte nennen diesen Krebs auch Lungenkarzinom oder Bronchialkarzinom

Das Bronchialkarzinom tritt hauptsächlich im Alter zwischen 50 und 70 Jahren auf. Das durchschnittliche Erkrankungsalter bei Diagnosestellung beträgt nur 69 Jahre.

Im Frühstadium von Lungenkrebs gibt es fast keine typischen Anzeichen und Symptome. Deshalb bleibt es oft lange unentdeckt. Bei Lungenröntgenuntersuchungen finden Ärzte häufig nur zufällig den Lungenkrebs, da der Patient an chronischem Husten, Atemnot oder Kurzatmigkeit leidet.



Abbildung 2

Abbildung 1 -

Für den Lungenkrebs ist mit weitem Abstand das ***Rauchen*** die Ursache. Ungefähr 90 Prozent aller Bronchialkarzinom-Patienten sind oder waren Raucher. Dabei hängt das Risiko vor allem davon ab, wie viel der Betroffene geraucht hat. Fachleute drücken dies anhand der Packungsmenge aus: Je mehr Zigaretten geraucht wurden, desto höher ist das Krebsrisiko.

Doch nicht nur selber zu rauchen macht krank, auch Passivrauchen erhöht das Risiko, an Lungenkrebs zu erkranken – und zwar um den Faktor 1,3 bis 2. Wer also das Rauchen aufgibt, kann das Risiko für Lungenkrebs deutlich senken. Vor allem für sich selbst, aber auch für seine Mitmenschen.

Wie viele Raucher sterben jährlich am Rauchen?

Rauchen ist das größte vermeidbare Gesundheitsrisiko weltweit. In Deutschland sterben jedes Jahr 110.000 bis 140.000 Menschen an den Folgen des Rauchens. Es gibt mehr als 6 Millionen Menschen auf der Welt. Infolgedessen verursacht Tabakkonsum 15 % der Todesfälle bei Männern und 7 % der Todesfälle bei Frauen.

Wie viel Prozent der Raucher bekommen Lungenkrebs?

Erkranken Raucher häufiger an Lungenkrebs als Nichtraucher? Männer, die rauchen, erkranken 30-mal häufiger an Lungenkrebs als Nichtraucher. Frauen, die rauchen, setzen sich einem 9-fachen Risiko aus. Mit anderen Worten, 91% der Lungenkrebsfälle bei Männern und 65% der Lungenkrebsfälle bei Frauen werden durch Rauchen verursacht. 2013 starben etwa 35.000 Menschen an Lungenkrebs, 91 % waren männliche Raucher, jedes Jahr starben etwa 31.000 Menschen an Lungenkrebs, der mit dem Rauchen in Zusammenhang steht.



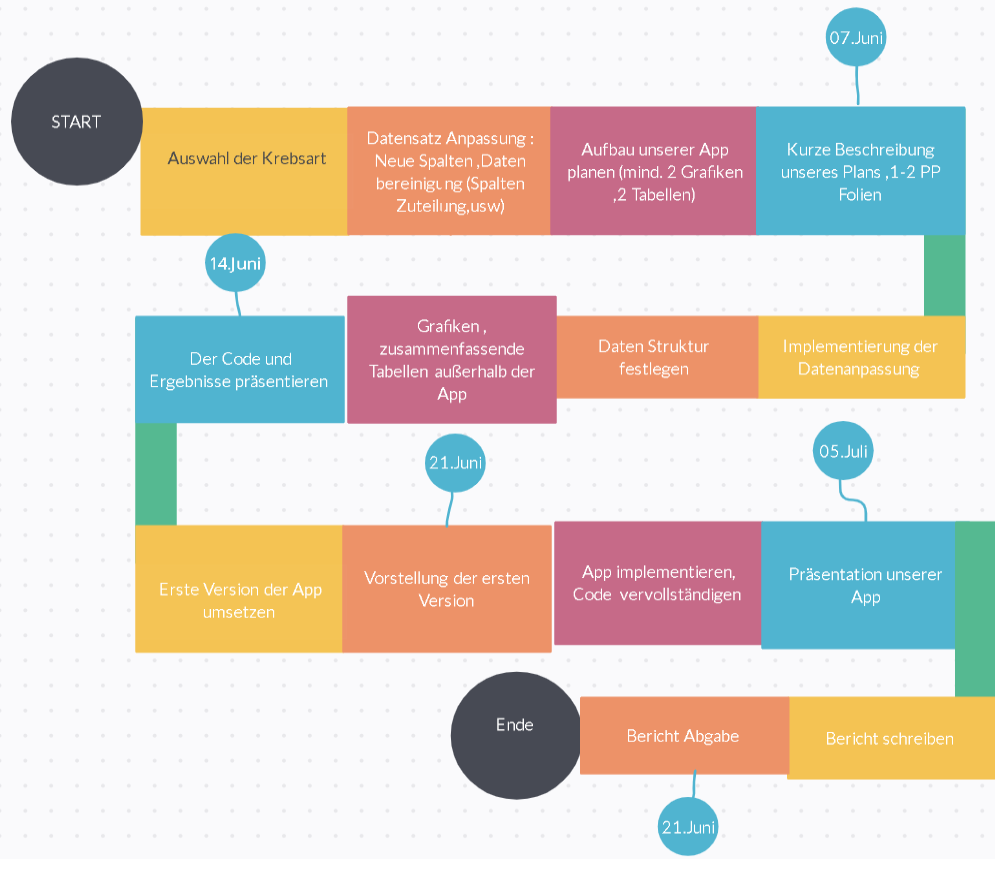


Abbildung 3 NICER – Neuerkrankungen; BFS - Sterbefälle

Unsere Aufgabe ist es nun mit einem Datensatz, die oben genannten Fakten bzgl. des Lungenkarzinoms bei ShinyApps zu visualisieren und zu testen.

**Planung des bearbeiteten Datensatzes und der App**

Zur Realisierung des gesamten Projekts haben wir uns im Team an diesem Zeitplan orientiert:



Als wir im Meeting den Datensatz näher betrachten hatten, fanden wir die Variable „Rauchen“ sehr interessant. Denn das Rauchen ist, nehmen wir an, sehr bekannt als die Ursache des Lungenkarzinoms. Das Rauchen kann hier sehr gut in Tabellen und Graphen verwendet werden. Desweiteren, fiel uns auf, dass „Geschlecht“ und „Rauchen“ indentische Werte vorweisten.

Um Ergebnisfehller bzw. Missverständnisse zu vermeiden und eindeutige Ergebnisse visualisieren zu können, haben wir Das Geschlecht von „0“ bzw. „1“ zu „m“ bzw „w“ verändert. Die Liste wurde neu sortiert. Die obligatirische „ID“ haben wir selbstverständlich an erster Stelle, denn so sieht die Liste professionell und auch für den Leser strukturiert aus. „Height und „Weight kommen nacheinander, das diese eine Gruppe aus den Körpermaßen bilden

Neu Hinzugefügte Variabeln

* Tumour\_size-dichotom
  + Der Tumour wurde hier mit T1 und T2 in zwei stufen eingeteilt. Ab einer Größe von 2cm hat der Betroffene die T2
* Altersgruppen
  + Durch Altergruppen lassen sich zusätzlich bessere Schlussfolgerungen ziehen, wie z.B bei den Graphen ( Säulendiagramm
* BMI
  + Ein erhöhter Cholesterinespiegel kann durch einer fettreichen Ernährung hervorgerufen werden, somit ist der BMI eine wichtige und eine aussagekräftige Variable.
* Chol\_dichotom
  + Der Cholesterinespiegel wurde mit niedrig, normal und hoch in drei Stufen eingeteilt. Ab einem Cholwert von 28 ist dieser „Normal“ und ab einem Cholwert von 29 ist dieser “Hoch“.
* BMI\_dichotom
  + Der BMI wurde mit „Untegewichtig“, „Normal“ und „Übergewichtig“ in drei Stufen eingeteilt. Unter einer BMI ist die betroffene Person untergewichtig, zwischen 18.5 und 25 normal und ab 25 übergewichtig

Planung von Grafiken und Tabellen

* Grafiken
  + Scatterplot

Streudiagramme oder Punktwolken können verwendet werden, um mögliche Beziehungen zwischen zwei Variabeln grafisch darzustellen. Dies ist oft hilfreich, wenn Sie diese Beziehung als Funktion beschreiben möchten. Geplant sind interaktive Scatterplots, in denen man die Variabeln beliebig ändern kann.

* + Boxplot

Boxplots, auch Box-Whisker-Plots oder deutsche Kastengrafik genannt, sind Diagramme, das die wichtigsten und zuverlässigsten Positionen und Streuungsmessungen übersichtlich darstellen kann. Dabei werden Minimum, unteres Quartil, Median, oberes Quartil und Maximum dargestellt. Geplant sind hier ebenfalls eine verstellbare bzw interaktive Darstellung mit Boxplots

* + Säulendiagramm

Mit Säulendiagrammen werden Datenreihen mittels Säulen grafisch darstellen. Dies ist sowohl mit absoluten als auch mit relativen Häufigkeiten möglich. Verstellbare bzw Interaktive Darstellung ist geplant.

* Tabellen
  + Kreuztabellen

In einer Kreuztabelle oder Kontingenztabelle können jeweilige absoluten Häufigkeiten verwendet werden, um statistische Merkmale oder Variablen anzuzeigen (normalerweise nominale oder geordnete Skalenmerkmale, da es normalerweise zu viele Werte gibt, um Merkmale zu messen). Dies ermöglicht die Anzeige und Auswertung von gemeinsamen Verteilungen oder gemeinsamen Häufigkeiten, Randverteilungen oder Randhäufigkeiten und bedingten Verteilungen.

Anschließend ist geplant, mittels Kreuztaballen, Abhängigkeitstest oder Homogenitätstest durchzuführen.

**Umsetzung des Datensatzes**

Die Umsetzung des Datensatzes erfolgte im Wesentlichen so wie es in der Planung vorgesehen war. Durch die Anpassung des Datensatzes wurden die Daten, die wir für unsere Krebsart untersuchen wollen, optimiert. Aus Gründen der Übersichtlichkeit wurde die Spalte ID an erster Stelle gesetzt. Zum besseren Verständnis der Einträge wurden in der Spalte Gender numerische Werte durch Strings ersetzt (0=männlich, 1=weiblich). Gleiches trifft für die smoking-Variable zu (0=nicht-raucher, 1=raucher). Weil wir die Werte der Spalte Diet nicht weiter für unsere App nutzten, wurde die gesamte Spalte entfernt. Da Height und Weight zusammenhängende Variablen sind haben wir beide Spalten mittles relocate nacheinander gesetzt und mit dem round-Befehl auf ganze Zahlen gerundet.

Für die geplanten statistischen Untersuchungen, Tabellen und Grafiken wurden neue Variablen in neuen Spalten angelegt: Die Patienten wurden in 4 verschiedenen Altersgruppen unterteilt. tumour\_size setzt sich als dichotomisierte Variable auf Grundlage der T-Achse der TNM Klassifikation (T1 & T2) zusammen. In einer weiteren neu angelegten Spalte findet man den BMI, errechnet aus der allgemeinen Formel mit Hilfe von den gerundeten Werten aus Grösse und Gewicht. Für fehlenden Werte (N.A.) aus der Cholesterin Spalte wurden Mittelwerte getrennt nach Geschlecht errechnet und in den fehlenden Eintragungen befüllt.

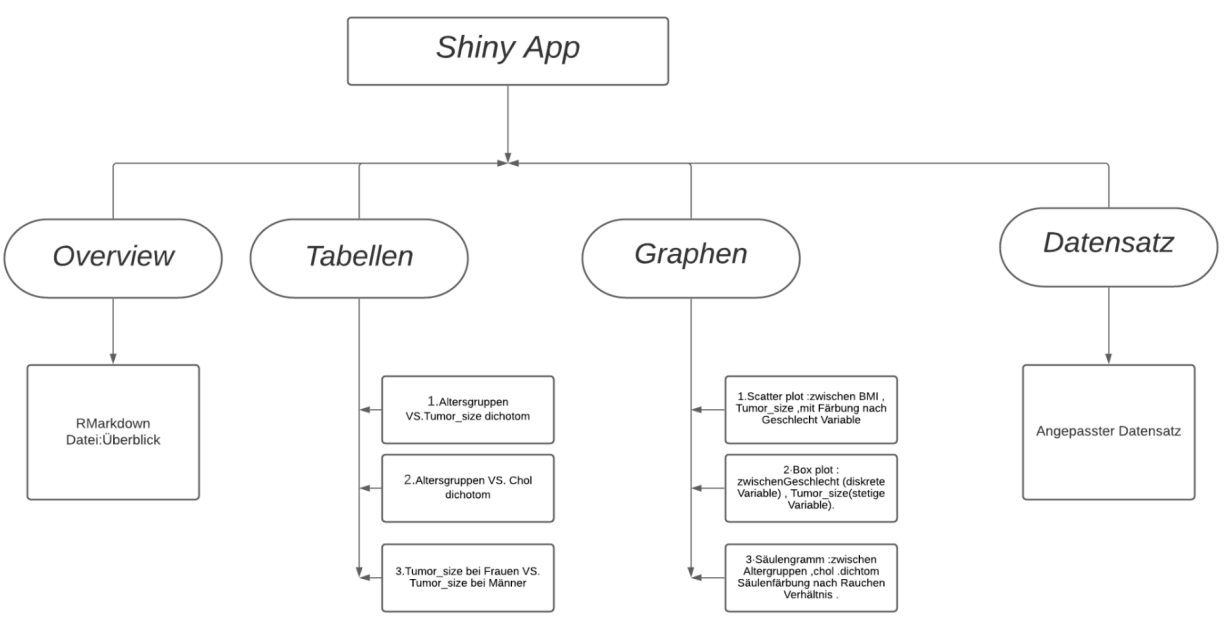
Abschließend sind im Datensatz dichotomisierte Werte für die Variablen Cholesterin (niedrig, normal, hoch) sowie BMI (Untergewicht, normal, Übergewicht) angehängt worden.

Nachdem nun sämtliche Anforderungen erfüllt waren, wurden alle Unterdatensätze in “data\_final” zusammen gefügt und auf dessen Grundlage nun die App erstellt wurde.

Für die Umsetzung wurden die R-Pakete **dplyr(Version 1.1.0)** und

**tidyr(Version 1.1.3)** genutzt.

**Umsetzung der App**



Durch die Tasks konnten wir uns vorstellen, wie die App aussehen wird und welche Bibliotheken im Rahmen der Umsetzung benötigt werden .

Die App wurde mit Hilfe von Shiny Bibliothek aufgebaut. Deswegen sollten wir zwei Skripte erstellen, einer für UI, einen anderen für den Server der App .

Unsere App besteht aus 4 Tabs in einem tabsetPanel,das auf einem mainPanel vom sidebarLayout dargestellt ist :

* Data Table
  + Auf diesem Tab wird der angepasste Datensatz gezeigt. Wie erwähnt, wurden die Anpassungsschritten wie geplant implementiert. Diese Tabelle wurde als eine Instanz vom Typ renderTable im output zurückgegeben, dann wird es im tableOutput auf dem User Interface gelesen und ausgegeben.
* Graphen
  + Scatter Plot :

Das Bibliothek ggplot2 bietet uns mit der Funktion geom\_point zwei stetige Variablen zu visaulisieren an. Auf dem wird die Korrelationstärke/-richtung mit Hilfe von Funktion geom\_smooth dargestellt.

Auswahl der Variablen: Durch selectInput wird eine Selectlist angezeigt, enthält als Auswahlen drei stetige Variablen: Cholesterinwerte, Tumorgrößen, BMI(Body Mass index).Der ineraktive Knopf („update“) erstellt den entsprechenden Graph mit der

Möglichkeit durch Radiobuttons eine diskrete Variable auszusuchen , um die Punkte im Scatterplot nach dieser Variabke zu färben .

* + Box Plot :

In demselben Bibiliothek besteht eine Funktion unter dem Name geom\_boxplot ,hier macht es sinn zu schauen, ob ein Unterschied in Tumorgrößen bei den Gruppen.

Die Einteilung der Patienten in den bestimmten Gruppen,wird entweder auf Rauchenverhältnis oder auf Geschlecht basiert. Hierfür haben wir Radiobuttons zum Aussuchen angewendet. Zusätzlich gibt es Möglichkeit ohne Einteilung mit None es darzustellen.

* + Säulendiagram :

Auf X-Achse werden die Altersgruppen versetzt. Unser Ziel war hier die relative und absolute Häufigkeit eine diskrete Variable zu schauen .Auf Y-Achse werden die absolute Häufigkeiten darzustellen. Die Säulenbefüllung wird durch diese diskrete zu zeichnen .

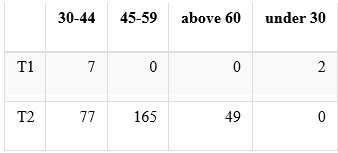
Der Auswahl erfolgt auch durch radioButtons Komponent .Die Variable sind Tumorgrößendichotom ,Cholesterindichotom und Geschlecht . Die Zeichnung eines Säulendiagrams wird mit Hilfe von Funktion geom\_bar durchgeführt .

Ein Beispiel : Im Fall nach Cholesterindichotom die Säulen zu befüllen, sehen wir die Altersgruppe zwischen45-59 , auf dieser Säule wurde im blauen Teil 29.7% , das bedeutet : 29.7% aller Patienten von allen Altersgruppen, sind im Alterbereich von 44-59 Jahre alt **und** haben normale Cholesterinwerte(Schnittsmenge) .

Alle Graphen wurden als Instanze von der Klasse renderPlot im output zurückgegeben ,und werden auf User Interface mit Hilfe von plotOutput gezeichnet.

* Tabellen

Die erste Tabelle ist eine Kreuztabelle zwischen den TNM klassifizierte Tumorgrößen und Altersgruppen. Mit Hilfe von der interngebauten Funktion xtabs. Um die Tabelle richtig formulieren zu können, sollte die Tabelle im Form vom Datentyp data.frame.matrix und anschließend im Instanz von der Klasse renderTable zurückgegeben werden. Auf der User Oberfläche im tableOutput wird dieses Instanz angezeigt .



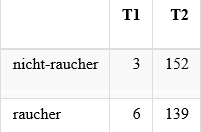
Hier gibt es keine Patienten über 45 Lebensjahr, deren Tumoren T1(<2 cm) nach TNM klassifiziert sind. Sowie bestehen keine jungen und an T2 Tumor erkrankten Patienten.

Für das Testen der Abhängigkeit zwischen Alter und Tumorgrößen, wird ein interaktives Button(„Abhängigkeit testen“) eingesetzt.

Teststatistik : Der Chi-Quadrad Test konnte nicht anwendet werden, weil mehr als 20% aller Zellen der Tabelle, kleiner als 5 sind. Deswegen sollte hier Exakter-Fisher Test angewerdet werden.

P-Wert ist viel kleiner als 5%. Der Zusammenhang zwischen Alter und Tumorgrößen kann nachgewiesen werden.

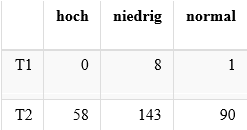
* Die zweite Tabelle ist eine Kreuztabelle zwischen Rauchen und Tumorgrößen:



All sinnvolles Maß wurde das Odds Ratio angewendet, da sie für den Vergleich der Risiken bei beiden Gruppen benötigt werden. Im Bibliothek epiR wird die Funktion epi.2by2 benutzt, um das Odds Ratio und Konfidenzintervall zu bestimmen.

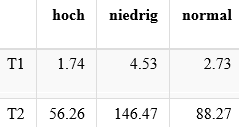
ODDS Ratio beträgt 0.457, das bedeutet, das Risiko bei Raucheren an großen Tumoren zu erkranken ist um 54 % höher als bei den Nicht-Raucheren.

* Die dritte Tabelle ist eine Kreuztabelle zwischen Cholesterinwerte und den Tumorgrößen



Aus der Tabelle können wir sehen, dass es keine Patienten gibt, die hohe Cholesterinwerte und große Tumoren haben. Hier ist festzustellen, ob die Tumorgrößen bei allen Cholesterin-Gruppen gleich sind.

Hier wollten wir untersuchen, ob die Cholesterinwerte bei beiden Tumorklassifikationen gleich sind. Deswegen sollten die Formel von erwarteten Häufigkeiten eingesetzt werden, um die Erfüllung der Voraussetung vom Homogenitätstest zu bestätigen.



P-Wert beträgt 0.058 ist größer als 0.05 %, d.h es spricht nichts gegen die Gleichheit der Tumorgrößen in allen Cholesteringruppen .

**About**

Das About Tab enthält sämtliche allgemeine Informationen, z.B warum wir Lungenkrebs ausgewählt haben, Namen von Teammitgleider, welche Packages wir im Rahmen der Projektarbeit benutzt haben.

Die geplannte R-Markdown Datei in der About section wurde durch einen herkömlichen Fließtext ersetzt.

Zur Umsetzung der App wurden folgende R-Pakete verwendet:

**Shiny (Version 1.6.0), tidyverse (Verson 1.3.1), epiR (Version 2.0.26), ggplot2 (Version 3.3.3)**

**Übersicht über die Eigenbeteiligung**

**Devrim:**

Durch das Projekt konnte ich einen umfassenden Eindruck gewinnen. Die in den Vorlesungen vermittelten Lerninhalte konnten bei der Ausarbeitung dieses Projektes angemessen angewendet werden und bestätigen den praxisnahen Bezug.

Zum Start der Projektarbeit haben wir die Aufgaben aufgeteilt. Anschließend haben wir nach weiteren Terminen unsere Codes gegenseitig präsentiert. Bei auftretenden Problemen wurden diese analysiert und ausdiskutiert. Die Projektgruppe hat sich mit den Mitgliedern abgestimmt, und die Probleme konnten gelöst werden.

Die Projektgruppe hat die gesamte Projektarbeit in drei großen Themen gegliedert. Ich war zuständig für die Grafiken, wobei bei Problemen, die auftauchten, andere Mitglieder mir geholfen haben. Das Gleiche gilt auch andersrum, da ich auch viel mit Erfolg versuchte zu den Problemlösungen beizutragen. Ich habe noch dazu beigetragen das Projekt etwas interaktiver zu gestalten (zb. Selectnput usw.), und die App mit Informationen (About) ergänzt.

Während des Projektes konnten verschiedene Eigenschaften, wie zum Beispiel, die Organisation von Besprechungen und Umgang mit plötzlich auftretenden Problemen erlernt werden.

**Benjamin:**

Nachdem am Projektstart die Aufgaben selbstständig und gerecht zugewiesen wurden, bestand mein Part darin, den Datensatz für unsere App anzupassen und in der App zu Implementieren. Konkret bedeutete dies, Spalten im Datensatz neu anzuordnen, Spalten zu entfernen, sie neu zu befüllen oder zu löschen. Zudem habe ich noch weiter Spalten hinzugefügt, die später (zB. bei zusammenhängenden Tabellen) relevant wurden. Sämtliche Fortschritte wurden auf das gemeinsame Repository auf github gepusht, sodass jeder auf dem aktuellen Stand des Projekts war.

Außerdem versuchte ich zur Problemlösung von anderen Teammitgliedern zur Stelle zu sein, egal ob es Probleme bei der Interaktion von Grafiken oder um die Implementierung von Buttons auf der grafischen Benutzeroberfläche ging. Dieses Projekt hat mir sehr dabei geholfen, meine Fähigkeiten als Teamplayer anzuwenden und zu vertiefen. Der Austausch im Team erfolgte sehr kommunikativ und reibungslos. Allumfassend lässt sich sagen, dass jeder im Team

**Kussi:**

Am Anfang habe ich die Aufgaben im Kontext der Anforderung in einer Word Datei (Tasks im Git-Repo) zusammengefasst. Diese Datei versuchte ausführlich wie möglich zu formulieren, damit jeder seinen Aufgabenbereich aussuchen kann. Darin enthalten, zu jedem Stichpunkt, (Tabs, Graphen, Tabellen, Anpassung des Datensatzes) alle Schritte, die im Rahmen der App-Implementierung benötigt werden z.B. Anpassungsschritte (fehlende Werte umgehen, welche neue Variable sollen eingefügt werden? Encoding, etc.), Graphen Typen (welche Variable (diskret, stetig) mit welcher Variablen eine Visualisierung Sinn macht). Auch habe ich die erwarteten Schlüsse bei den Teststatistiken gefasst, was wir zu jeder Tabelle testen wollen (Unabhängigkeit). Danach half ich die andere Mitglieder beim Klonen des Git-Repo, sowie bei den Hauptfunktionen (commit , push..) anzuleiten. Mein Aufgabebereich war die zusammenfassenden Tabellen zu erstellen, mit geeigneten Teststatistiken zu kombinieren, letztendlich die Ergebnisse zu interpretieren.

Daneben ist den passenden Maßen anzuwenden, um sinnvolle Schlüsse aus den Tabellen zu ziehen. Außerdem unterstützte ich die Mitglieder bei der Anpassung des Datensatz und bei der Erstellung der Graphen. Zusätzlich zu den Problemlösungen zu finden, war berücksichtigt alle Anforderung fehlerlos zu implementieren. Ich versuchte immer ersten Schritten mit Beispielen zu schaffen, um die anderen identischen Schritte zu verfolgen, und erfolgreich mit anderen Stücken zusammenzusetzen. Am Ende fügte ich die Interaktionen an den Graphen an, wobei unsere App breite Visualisierungsmöglichkeiten anbieten kann, z.B. die Befüllung der Säulen beim Graph 3 nach einer durch User gewählten diskreten Variable mit den Prozentanteile von allen Stichproben durchführen.

**Fazit**

Die geplanten Aufgaben waren möglich fehlerlos zu implementieren, da wir möglichst zu jeder machbaren Unterpunkten Stichpunkte eingeteilt haben, um die Herausforderungen bei unzureichenden praktischen Kenntnissen zu beseitigen und früh zu erkennen.

Den Schwergrad des Projektes würden wir auf 6 von 10 setzen, da wir genügend Hilfe von den Vorlesungsfolien bekommen haben. Auch der Coding-Anteil machte uns Spaß und die Problemlösungen motivierten uns sehr.

Durch dieses Projekt vertieften wir unsere Kenntnisse in R und lernten die breiten Nutzungsmöglichkeiten des Shiny- Dashboards. Von diesem Projekt nehmen wir in R, vielfältige Erfahrungen bei Bearbeitung der Problemlösungen mit.

Der gemeinsame Erfolg beim Einhalten der Deadline und beim Aufbau unserer App motivierte uns sehr viel, um in Zukunft weitere Tools in R zu lernen. Außerdem nehmen wir an, dass wir bei einer größeren Datensatz mit mehreren Spalten, noch mehr Analysentechniken umsetzen könnten, die wir im Rahmen der Statistikmodulen gelernt haben.

**Quellenverzeichnis:**

Abbildung 1 – https://www.bma.org.uk/news-and-opinion/follow-the-evidence-improving-public-health

Abbildung 2-https://www.medica.de/de/News/Thema\_des\_Monats/%C3%84ltere\_Themen\_des\_Monats/Themen\_des\_Monats\_2018/Bildgebung\_der\_Lunge/Bildgebung\_der\_Lunge\_%E2%80%93\_f%C3%BCr\_eine\_gesunde\_Atmung

Abbildung 3- https://www.nicer.org/

https://www.massmatics.de/merkzettel/#!963:Scatterplot

https://studyflix.de/statistik/boxplot-1044

<https://studyflix.de/statistik/histogramm-1711>