STAT 6080 Data Technologies

Project Description

by

Jürgen Symanzik

Date: February 26, 2023

Final Due Date: Saturday, April 29, 2023, 11:59pm (by e–mail)

UTAH STATE UNIVERSITY

Logan, UT

Spring 2023

Contents

1 Background Information

In this Data Technologies (DT) course project, you can scrape data in html, XML, or pdf format from the web, or you can do some extensive data manipulation via regular expressions or using the functionality from *tidyverse* or work with OCR. The data should not be readily available in any formats that are directly supported by R or Microsoft Excel (such as rda, txt, csv, xls, xlsx, etc.). You can also do a simulation study as part of the project. This project can be closely related to your MS or PhD research or to an outside job or project. However, this should be a side–project with respect to your overall MS or PhD research and should not be used as a chapter of your MS report or dissertation. You also should not be paid by any source for conducting this project.

The project is a group project with exactly three students in a group!

In a few cases in past years, interested students continued to work on this project after the end of this course and eventually were able to transform this project into a conference presentation and an accompanying proceedings paper. Other students continued to work on a second stage of their projects in one of my follow–up courses such as *Statistical Visualization II*, *Applied Spatial Statistics*, or *Data Technologies*. If this is of interest to you, let's further discuss this once the course has ended.

This project consists of multiple stages, outlined in the following section.

2 Tasks and Timeline

- **Preliminary Discussion of Project Proposal:** If you have some ideas for a project that meets the overall ideas outlined above, I would like to hear your suggestions! If you don't have any ideas (or no MS or no PhD topic yet and also no suitable outside project), I have some possible suggestions for you. You have to meet with me on Tuesday 2/21/23 for this preliminary discussion.

- **Written Project Proposal:** Based on the preliminary discussion of your project proposal, prepare a full two–page project proposal. This is kind of a road map to a successful completion of your project. You must indicate which data set(s) or web pages you are going to use and which existing R packages (and functions) you are going to use. Be specific how you are going to use/apply/modify/extend the existing functionality. It must become clear what already exists and what needs to be implemented by you. Be realistic and only suggest what can be done over a six–week period. Cite and list supporting references (at least any required R packages needed for your project). Small graphics, diagrams, or sketches are fine to support your proposal. R code is not needed at this time.

  Also provide some information which group member is primarily in charge for which tasks of the project, e.g., for the R code for each of the main software components, as well as for the written project proposal, the final project presentation slides, and the

written project report. Different students must be in charge for the different tasks. In particular, three different students have to be in charge for the written project proposal, the final project presentation slides, and the written project report. The student who is in charge of a certain task has to set deadlines for the other group members, collect materials from the other group members, and finalize that task prior to the deadline (and submit the deliverable to me by the deadline). Apparently, as this is a group project, all group members are encouraged to contribute to each task, check for correctness and plausibility, and proofread the deliverables (such as the proposal, the slides, the final report, etc.).

Deliverables: Please e–mail your written project proposal by Wednesday 3/1/23, 11:59pm. As soon as you get my approval, you should start working on your project. If you submit your proposal early, I will try to provide feedback as soon as possible so you will have a few extra days to work on your project.

- **Weekly Progress Reports:** Your group has to give a short 10 min progress report each week, usually on Thursdays, specifically on 3/23/23, 3/30/23, and 4/6/23. In these short progress reports, you should present what you have done since the previous progress report and what your next planned steps are. The other students in class and I may ask questions and make suggestions. We all should agree what the next steps and priorities are. In the first weekly report on 3/23/23, you should present an overview of your planned project to the entire class and provide some initial glances at the underlying data and show sketches of some visualizations (where appropriate). Some slides with the names of the group members, the overall topic of the project, and the work presented in a weekly progress report are helpful for the audience.

  Two students are in charge of the progress report in a given week. Each student **must** be a presenter twice for the weekly progress reports. These are in–class presentations where the presenters are expected to present in class (except in case of an emergency).

- **Final Project Presentation:** You have to prepare a 20 min presentation of your work. This presentation should contain an introduction into the underlying problem, an overview of the data, methods, and software used, and a brief summary of the supporting literature (where appropriate). The main part should be a summary of the computational methods and R packages you have used, and obviously a presentation of your results, including some numerical and graphical summaries. End with a discussion/conclusion/outlook on things you would have liked to do (from your written project proposal), but were not able to do given the time limits of this project and possible limitations of the data you used. It is not necessary to come up with final answers to all initial questions you had planned to answer. Also include a slide with a short list of references (related published research, data sources, and references related to main R packages). Use my LaTeX Beamer slides (in Presentation.zip) as a template for this presentation. Examples of past project

presentations can be located in PastProjects.zip.

Your group will present your final project presentation to the other students from this course and other people from the department interested in the work conducted in this class (e.g., from one of my other courses). Imagine that this is a contributed session at the Joint Statistical Meetings (JSM), something many of you likely will experience at some point in the future (or already have experienced in the recent past).

On Thursday 4/13/23, your group has to give this 20 min presentation of your final results via Zoom (as this will take place during one of our Backup lecture slots). This date may shift to the following week (Thursday 4/20/23) if necessary.

While there could be a "narrator" for the overall presentation, each student most contribute to it and report about the main part to which he/she contributed. All students must be able to answer general questions about the project and specific questions about their main part.

- **Written Project Report:** Summarize your project in a final written project report that resembles a first short (six–page) draft of a proceedings paper for the Joint Statistical Meetings (JSM). There must be a title, abstract, and keywords. Main sections are the introduction, methods (describing the data and previously existing methods and software you used), a section describing your results of this project, and a discussion/conclusion/outlook (on future work) section, followed by the reference list. Also, include your R code in the appendix. The appendix does not count towards the six–page limit.

  On Friday 4/28/23 at 11:59pm, your six–page final written project report is due. **This date is fixed!** Likely, you have to make a careful decision what to include in your written report. In a presentation, we can often present a lot of additional information that won't make it into the final written report due to space (i.e., page) limitations.

  Use my LaTeX template (in ProceedingsPaper.zip) as a template for this final report. Examples of past written project reports can also be located in PastProjects.zip.

- **Contributions to the Project:** On Saturday 4/29/23 at 11:59pm, your e–mail with your contributions and the percent–wise breakdown is due. See below for details.

3 Grading

- Preliminary discussion of the project proposal: 2% of your total score for this project.

- Written project proposal: 7%.

- Three weekly progress reports in class: 9% (i.e., each 3%).
- Final project presentation in class on 4/13/23: 40%.
- Final written project report due on 4/28/23: 40%.
- e–mail with the breakdown of your contributions on 4/29/23: 2%.

4 General Instructions

- Generalize from the instructions from previous homeworks in this course! All programming must be done in R (possibly with a combination of other external programming languages and tools if needed for your project). You cannot manually manipulate files and data. Your code should be reproducible/reusable, e.g., when new, corrected, or updated data sets become available.
- Your written project proposal and the final project report must be prepared with LaTeX, sweave, and/or knitr. Your presentation has to be prepared with the LaTeX–beamer package. Note that LaTeX–beamer is a high–level LaTeX–based approach that replaces PowerPoint presentations. Please submit all your source and resulting output files (Rnw, pdf, tex, figures, local data files, etc.) for your final version of the project paper and your presentation. For the written project proposal, you only have to submit the pdf file via e-mail.
- Before you submit your files, make sure that everything works and translates on a computer other than your own computer! Your documents must be fully reproducible on a different computer, i.e., I must be able to re–translate your files without encountering any errors. If you haven't done so at the start of the semester, update R and all of your R packages to a recent version of R such as 4.2.2 (use the *installr* R package to quickly update to a new version of R).
- I will provide several templates, such as those for the JSM proceedings and a JSM presentation, all in LaTeX & LaTeX–beamer format (and also the underlying source files). These will be placed in *Discussions* under the *Project* topic in Canvas.
- It is up to you how to split the individual tasks, but different students must be in charge for the different tasks (as outlined in your written project proposal). Each group member **must** present twice during the weekly progress reports. Each group member **must** contribute to the final presentation.
- It is not expected that each group member contributes exactly the same amount of time to a group project. However, no single group member is allowed to contribute more than 45% of the overall work for this project. Each group member that contributes less than 20% of the overall work to this project will get individual point

deductions. You have to provide me with an estimated percentage of everyone's contributions to this project and the kind of contributions everyone made.

- Think of efficient high–end visualizations of your final results (where appropriate). I would like to see at least two different meaningful high–end visualizations such as maps, heat maps, timeline plots, scatterplot matrices, or parallel coordinate plots to mention only a few (as appropriate for your data) in your written project report and possibly more in your final presentation. Interactive graphics (e.g., via Shiny) often are beneficial for a course project in *Statistical Visualization II* and *Applied Spatial Statistics*, but usually are not needed for a course project in *Data Technologies*.

- A final six–page written project report is due on Friday 4/28/23 at 11:59pm. This has to follow the JSM formatting requirements from the American Statistical Association (ASA). These six pages must contain everything from an abstract, keywords, the main sections of your report, all figures and tables, and the references. I will provide a template. In addition, you have to arrange your R files in a meaningful way as in Appendix A of your main report. There is no page limit for this appendix. LaTeX commands such as *verbatiminput* exist and allow you to create your R code independently and only include it into a document at the very last stage. You have to turn in your final pdf file and all source files.

- Please e–mail your breakdown of everybody's percent–wise contribution to the project and the individual responsibilities by Saturday 4/29/23 at 11:59pm. You should not cc your other group members on this e–mail and just send it to me. Ideally, each group member should have contributed to the project as outlined in the written project proposal, but in practice, there are often differences, in particular if a students drops out of the project. See Appendix B for such a breakdown (but without any percentages). This is common for publications in the medical field. Apparently, I am "J.S." in the example.

- When you reach out to me via e–mail, always cc your other group members (with the exception of the percent–wise contributions mentioned above or any other personal questions).

- **Whenever you have questions or need clarifications, talk to me in person, via e–mail, or via Zoom or Skype. Good luck!**

Appendix A Rnw File for the Project

This appendix contains the Rnw file for this project description:

```
\documentclass[12pt,letterpaper,final]{article}

\usepackage{graphicx}
\usepackage{url}
\usepackage{hyperref}
\usepackage{verbatim}
\usepackage[title,titletoc,toc]{appendix}
\usepackage{wasysym}

\renewcommand{\topfraction}{1.0}
\renewcommand{\bottomfraction}{1.0}
\renewcommand{\textfraction}{0.0}
\renewcommand{\floatpagefraction}{1.0}
\renewcommand{\dbltopfraction}{1.0}

\textwidth 16cm
\textheight 22cm
\voffset -0.5in
\hoffset -0.5in

\parindent0pt
\setlength{\parskip}{1ex plus 0.5ex minus 0.2ex}


\begin{document}

\begin{titlepage}

\begin{center}
{\large STAT 6080 Data Technologies} \\[4cm]
%{\large STAT 6560 Statistical Visualization II} \\[4cm]

{\LARGE \bf Project Description} \\[1cm]
by \\[0.5cm]
{\bf J\"urgen Symanzik} \\[2.5cm]
{\bf Date:} \today \\[2cm]
{\bf Final Due Date:} Saturday, April 29, 2023, 11:59pm (by e--mail) \\[2cm]

UTAH STATE UNIVERSITY \\[0.5cm]
Logan, UT \\[0.5cm]
Spring 2023 \\[0.5cm]
\end{center}

\thispagestyle{empty}
\vfill
\end{titlepage}


\newpage


\pagenumbering{roman}

\tableofcontents


\newpage
```

```
%\listoftables
%\addcontentsline{toc}{section}{List of Tables}
%
%\newpage
%
%\listoffigures
%\addcontentsline{toc}{section}{List of Figures}
%
%\newpage


\pagenumbering{arabic}


\section{Background Information}

In this
%Statistical Visualization II (Stat Viz II)
Data Technologies (DT)
course project,
%you have to use or extend a Stat Viz method we have discussed in class
%or apply a different visualization method not discussed in class to a data set.
you can scrape data in html, XML, or pdf format from the web, or you can do some
extensive data manipulation via regular expressions or using the
functionality from {\it tidyverse} or work with OCR. The data should not be readily
available in any formats that are directly supported by R or Microsoft Excel
(such as rda, txt, csv, xls, xlsx, etc.).
You can also do a simulation study as part of the project.
%
This project can be closely related to your MS or PhD research
or to an outside job or project. However,
this should be a side--project with respect to your overall MS or PhD research
and should not be used as a chapter of your MS report or dissertation.
You also should not be paid by any source for conducting this project.

{\bf The project is a group project with exactly three students in a group!}

In a few cases in past years, interested students continued
to work on this project after the end of this course and eventually
were able to transform this project into a conference presentation and
an accompanying proceedings paper. Other students continued to work on
a second stage of
their projects in one of my follow--up courses such as
{\it Statistical Visualization II}, {\it Applied Spatial Statistics},
or {\it Data Technologies}.
If this is of interest to you,
let's further discuss this once the course has ended.

This project consists of multiple stages, outlined in the following
section.


\section{Tasks and Timeline}

\begin{itemize}
\item {\bf Preliminary Discussion of Project Proposal:}
If you have some ideas for a
project that meets the overall ideas outlined above, I would like to hear
your suggestions! If you don't have any ideas
(or no MS or no PhD topic yet and also no suitable outside project),
I have some possible suggestions for you. You have to meet with me
on Tuesday 2/21/23 for this preliminary
discussion.
```

```
\item {\bf Written Project Proposal:}
Based on the preliminary discussion
of your project proposal, prepare a full two--page project proposal.
This is kind of a road map to a successful completion of your project.
You must indicate which data set(s) or web pages you are going to use and which
existing R packages (and functions) you are going to use.
Be specific how you are going to use/apply/modify/extend the existing
functionality. It must become clear what already exists and
what needs to be implemented by you. Be realistic and only
suggest what can be done over a six--week period.
Cite and list supporting references (at least any required R packages
needed for your project). Small graphics, diagrams, or sketches are fine to support
your proposal. R code is not needed at this time.

Also provide some information which group member is primarily in charge for
which tasks of the project, e.g., for the R code for each of the main software components,
as well as for the written project proposal, the final project presentation slides,
and the written project report. Different students must be in charge for the
different tasks. In particular, three different students have to be in charge
for the written project proposal, the final project presentation slides,
and the written project report. The student who is in charge of a certain task has to set deadlines
for the other group members, collect materials from the other group members,
and finalize that task prior to the deadline (and submit the deliverable to me
by the deadline). Apparently, as this is a group project, all group members
are encouraged to contribute to each task, check for correctness and
plausibility, and proofread the deliverables (such as the proposal,
the slides, the final report, etc.).

{\bf Deliverables:} Please e--mail  your written project proposal
by Wednesday 3/1/23, 11:59pm. As soon as you get my approval,
you should start working on your project.
If you submit your proposal early, I will try to provide feedback
as soon as possible so you will have a few extra days to work
on your project.
%I will read the proposals in a first--in first--out (FIFO) order.


\item {\bf Weekly Progress Reports:}
Your group has to give a short 10~min progress report each week,
usually on Thursdays, specifically on 3/23/23, 3/30/23, and 4/6/23.
In these short progress reports,
you should present what you have done since the previous progress report and
what your next planned steps are. The other students in class and I may ask questions and
make suggestions. We all should agree what the next steps and priorities are.
In the first weekly report on 3/23/23, you should present an overview of your
planned project to the entire class and provide some initial glances at the
underlying data and show sketches of some visualizations (where appropriate).
Some slides with the names of the group members, the overall topic of the project, and the work
presented in a weekly progress report are helpful for the audience.

{\bf Two students} are in charge of the progress report in a given week.
Each student {\bf must} be a presenter twice for the weekly progress reports.
These are in--class presentations where the presenters are expected
to present in class (except in case of an emergency).


\item {\bf Final Project Presentation:}
You have to prepare a 20~min presentation of your work.
This presentation should contain an introduction into the underlying problem,
an overview of the data, methods, and software used,
and a brief summary of the supporting literature (where appropriate).
The main part should be a summary of the
computational methods and R packages you have used, and obviously a presentation of your results,
including some numerical and graphical summaries.
End with a discussion/conclusion/outlook on things you would have liked to do (from your written
```

```
project proposal),
but were not able to do given the time limits of this project and possible limitations of the data you used.
It is not necessary to come up with final answers to all initial questions you had planned to answer.
Also include a slide with a short list of references
(related published research, data sources, and references related to main R packages).
Use my \LaTeX\ Beamer slides (in Presentation.zip) as a template for this presentation.
Examples of past project presentations can be located in PastProjects.zip.

Your group will present your final project presentation to the other
students from this course and other people from the department interested
in the work conducted in this class (e.g., from one of my other courses). Imagine that this is
a contributed session at the Joint Statistical Meetings (JSM), something many of you
likely will experience at some point in the future (or already have experienced in the
recent past).

On Thursday 4/13/23, your group has to give this 20~min presentation of your final results via Zoom
(as this will take place during one of our Backup lecture slots).
This date may shift to the following week (Thursday 4/20/23) if necessary.

While there could be a ``narrator'' for the overall presentation, each student
most contribute to it and report about the main part to which he/she contributed.
All students must be able to answer general questions about the project
and specific questions about their main part.


\item {\bf Written Project Report:}
Summarize your project in a final written project report that
resembles a first short (six--page) draft of a proceedings paper
for the Joint Statistical Meetings (JSM). There must be a
title, abstract, and keywords. Main sections are the
introduction, methods (describing the data and previously
existing methods and software you used), a section
describing your results
of this project, and a discussion/conclusion/outlook (on future work)
section, followed by the reference list. Also, include your
R code in the appendix. The appendix does not count
towards the six--page limit.

On Friday 4/28/23 at 11:59pm, your six--page final written project report is due.
{\bf This date is fixed!} Likely, you have
to make a careful decision what to include in your written report. In a presentation,
we can often present a lot of additional information that won't make it into the
final written report due to space (i.e., page) limitations.

Use my \LaTeX\ template (in ProceedingsPaper.zip) as a template for this final report.
Examples of past written project reports can also be located in PastProjects.zip.


\item {\bf Contributions to the Project:}
On Saturday 4/29/23 at 11:59pm, your e--mail with your contributions and the
percent--wise breakdown is due. See below for details.

\end{itemize}


\section{Grading}

\begin{itemize}
\item Preliminary discussion of the project proposal: 2\% of your total score for this project.
\item Written project proposal: 7\%.
\item Three weekly progress reports in class: 9\% (i.e., each 3\%).
\item Final project presentation in class on 4/13/23: 40\%.
\item Final written project report due on 4/28/23: 40\%.
\item e--mail with the breakdown of your contributions on 4/29/23: 2\%.
\end{itemize}
```

```
\section{General Instructions}

\begin{itemize}
\item Generalize from the instructions from previous homeworks in this course!
All programming must be done in R (possibly with a combination
of other external programming languages and tools if needed for your project).
You cannot manually manipulate files and data.
Your code should be reproducible/reusable, e.g., when new, corrected, or updated
data sets become available.

\item Your written project proposal
and the final project report must be prepared with \LaTeX,
sweave, and/or knitr. Your presentation has to be
prepared with the \LaTeX--beamer package.
Note that \LaTeX--beamer is a high--level \LaTeX--based approach
that replaces PowerPoint presentations.
Please submit all your source and resulting output
files (Rnw, pdf,  tex, figures, local data files, etc.)
for your final version of the project paper and your presentation.
For the written project proposal,
you only have to submit the pdf file via e-mail.

\item Before you submit your files, make sure that everything works and translates
on a computer other than your own computer!
Your documents must be fully reproducible on
a different computer, i.e., I must be able to re--translate
your files without encountering any errors.
If you haven't done so at the start of the semester, update
R and all of your R packages to a recent version of R
such as 4.2.2
(use the {\it installr} R package to quickly update to
a new version of R).

\item I will provide several templates, such as those
for the JSM proceedings and a JSM presentation,
all in \LaTeX\ \& \LaTeX--beamer format (and also the underlying source files).
These will be placed in {\it Discussions} under the {\it Project} topic in Canvas.

\item It is up to you how to split the individual tasks,
but different students must be in charge for the different tasks
(as outlined in your written project proposal).
Each group member {\bf must} present twice during the weekly progress reports.
Each group member {\bf must} contribute to the final presentation.

\item It is not expected that each group member contributes exactly the same amount
of time to a group project. However, no single group member is allowed to
contribute more than 45\% of the overall work for this project. Each group
member that contributes less than 20\% of the overall work to this project
will get individual point deductions. You have to provide me with
an estimated percentage of everyone's contributions to this project
and the kind of contributions everyone made.

\item Think of efficient high--end visualizations of your final results (where appropriate).
I would like to see at least two different meaningful high--end visualizations
such as maps, heat maps, timeline plots, scatterplot matrices, or parallel coordinate plots
to mention only a few (as appropriate for your data)
in your written project report and possibly more in your final presentation.
Interactive graphics (e.g., via Shiny) often are beneficial for a course project
in {\it Statistical Visualization II} and {\it Applied Spatial Statistics},
but usually are not needed for a course project in {\it Data Technologies}.

\item A final six--page written project report is due on Friday 4/28/23 at 11:59pm. This has
```

```
to follow the JSM formatting requirements from the American Statistical Association (ASA).
These six pages must contain everything from an abstract, keywords, the main
sections of your report, all figures and tables, and the references.
I will provide a template. In addition, you have to arrange your R files in a
meaningful way as in Appendix~\ref{AppendixWithRCode} of your main report. There is no page limit
for this appendix. \LaTeX\ commands such as {\it verbatiminput} exist and allow you to create your
R code independently and only include it into a document at the very last stage.
You have to turn in your final pdf file and all source files.

\item Please e--mail your breakdown of
everybody's percent--wise contribution to
the project and the individual responsibilities by Saturday 4/29/23 at 11:59pm.
You should not cc your other group
members on this e--mail and just send it to me.
Ideally, each group member should have contributed to the project as
outlined in the written project proposal, but in practice, there are often
differences, in particular if a students drops out of the project.
See Appendix~\ref{Contributions} for such a breakdown (but without any percentages).
This is common for publications in the medical field. Apparently,
I am ``J.S.'' in the example.

\item When you reach out to me via e--mail, always cc your other group members (with the
exception of the percent--wise contributions mentioned above or any other personal questions).

\item {\bf Whenever you have questions or need clarifications, talk to me in person,
via e--mail, or via Zoom or Skype. Good luck!}

\end{itemize}


\newpage


\begin{appendices}


\section{Rnw File for the Project}\label{AppendixWithRCode}

This appendix contains the Rnw file for this project description:

{
\scriptsize
\verbatiminput{DTProject.Rnw}
}


\newpage


\section{Author Contributions}\label{Contributions}

G.F. conceived the research, wrote the original manuscript, the
revised manuscript, and the response letter, and addressed the
reviewer comments; X.D. created the computer code, performed
the data analyses and created the figures; J.S. fine-tuned the figures
and verified the reproducibility of the results; S.B. performed
the STRUCTURE analysis and verified the biological interpretations;
all authors participated in discussions, read and revised the different
versions of the manuscript and the response letter, and agreed
to the submission.

\end{appendices}

\end{document}
```

Appendix B Author Contributions

G.F. conceived the research, wrote the original manuscript, the revised manuscript, and the response letter, and addressed the reviewer comments; X.D. created the computer code, performed the data analyses and created the figures; J.S. fine-tuned the figures and verified the reproducibility of the results; S.B. performed the STRUCTURE analysis and verified the biological interpretations; all authors participated in discussions, read and revised the different versions of the manuscript and the response letter, and agreed to the submission.