STAT 6080 Data Technologies

Project Description

by

Jürgen Symanzik

Date: October 29, 2021

Final Due Date: Friday, December 10, 2021, 11:59pm (by e–mail)

UTAH STATE UNIVERSITY

Logan, UT

Fall 2021

Contents

1 General Instructions

Your project is related to weekly advertisements from stores in the United States. These could originate from local food stores / supermarkets in Utah (such as Smith's, Macey's, Lee's Marketplace, etc.) or from across the US (archived at `https://weekly-ads.us/`, dating back to 2018). You could also work with ads from national general merchandise retailers (such as WalMart, Target, ShopKo, etc.), electronic store chains (such as Best Buy, etc.), home improvement store chains (such as Lowe's, Home Depot, Menards, etc.), or others.

The overall question is: How did weekly advertisements change after the outbreak of the Coronavirus pandemic in March 2020 and the recent problems with the supply chain shortages in the second half of 2021, i.e., how did these events affect what is listed in weekly store advertisements? For example, did toilet paper disappear from weekly ads in the middle of 2020 and did bananas disappear from weekly ads in the fall of 2021?

You have to work in groups of three to four students on this project. Each group has to decide on its own local or national chain. **Once your group has decided on a chain, claim your chain as quickly as possible to me via e–mail and cc the other groups on your e–mail.** No other group can claim the same chain thereafter.

There are three groups overall. Members of Groups 1 and 2 indicated that they would prefer to meet with other group members in person. Members of Group 3 indicated that they would prefer to meet via Zoom. Ultimately, it is up to each group to decide when (and how) to meet. Clearly, Groups 1 and 2 can meet via Zoom as well, e.g., on weekends.

I consider course projects to be generally open to input from others. All students are expected to frequently communicate and closely work together on overlapping questions, even across groups if that turns out to be helpful. Moreover, this project is open to comments and suggestions from the other students in class as well. So, you also have to consider suggestions made in class or in Canvas by the other students from this course. There are a few specific requirements and a timeline you have to follow. More details in the sections below.

2 Possible Research Questions

The overall goal of weekly advertisements is to attract shoppers to a store so that they purchase other non–sale items as well. Since the start of the Coronavirus pandemic outbreak in March 2020, our shopping behavior has changed considerably, e.g., far more online shopping, home deliveries, or pickups on the outside of a store. Moreover, there were supply shortages due to the pandemic, and more recently, due to global supply chain shortages. How did all of this affect what is advertised in weekly ads (and at which price)?

Here is a list of possible questions of interest to store managers and customers that are related to weekly ads. This list easily could be extended. You should select weekly ad

data and then provide answers to some of these questions. It is not expected that each group answers all possible questions.

- What should be placed on page 1 of a weekly ad?
- What do competitors put on sale?
- What are the prices of the advertised articles over time?
- How many items are listed in an ad — and how many on page 1? Did this change over time?
- What is the median price across all items in the ad / on page 1? Did this change over time?
- What are the different types of deals (33% off, buy 6 — save 60c on each, buy 1 — get one free or one half off, $xx off, etc.)?
- What type of information is provided to the customer, e.g., fixed price amount in the ad vs. amount of savings in the ad?
- What is the basis of the discount, e.g., on sale by unit (e.g., pound) vs. on sale by each?
- Are there any extras, e.g., buy more, add points (e.g., for fuel, ...)?
- What is the seasonality of advertised articles?
- Global min / max for advertised items. When/where was this item cheapest / most expensive over time across all locations?
- How frequently are items on sale? Is there a certain pattern, e.g., every 4 or 5 weeks?
- Are prices for advertised articles always the same, e.g., buy 3 for $10 vs. $3.99 each?
- How did Covid affect ad prices (e.g., shortages, store closures / reduced hours, etc.)?
- How did Covid affect which items are listed in ads?
- Are there visible effects of supply chain shortages on items in ads during the past 2 or 3 months (due to backlogs in US ports, truck driver shortages, etc.)?
- Additional question you can think of or that are suggested by other students from this course.

3 Specific Instructions

These are the main components for this project:

- All your work most be done in R. You cannot manually manipulate files and data. Your code should be reproducible/reusable, in particular for forthcoming weekly ads that will be released while you are working on the project
- Identify your chain. Then download their weekly ads in digital format. Do this over an extended period of time with random time intervals in between, e.g., a few minutes. Possibly do this from different computers, e.g., all 2021 ads from computer 1, all 2020 ads from computer 2, etc. You do not want to get blocked from further accesses because the ad server thinks your ad downloads are a malicious attack. Store each of the ads you downloaded locally on your harddrive, or even better, in a cloud drive such as Box, Dropbox, or Google Drive that can be shared among all group members.
- Depending on the file format of the weekly ads, you may have to use different approaches how to extract relevant product and price information. For pdf files, it should be straight forward to extract relevant information precisely and quickly. For jpg, png, or other image files, OCR will be a major component. Note that the `WeeklyAdOutline.R` file discussed in class should serve as a starting point and not as the end point to extract data from image files. You have to see how you can best clean an ad page from background noise (and images), make use of the bounding box of a text, or see whether the reported confidence information for a text could be used to filter out text that is not very reliable.
- Regular expressions likely will play a major role to extract relevant data and clean up the OCR results. But, do not forget to proofread a sample of weekly ads. The group member who initially comes up with the regular expressions is equally important as the group member who manually (i.e., visually) compares the extracted OCR results with the actual ad pages and comes up with shortfalls, omissions, and common interpretation problems of the OCR functionality. Apparently, this information should be used to optimze and fine–tune the regular expressions. Then, another visual assessment is necessary (and so on). Once you feel comfortable that you get meaningful results for this subset of ads, apply your methods to all ads. Even then, check for unexpected results and outcomes, e.g., if the chain unexpectedly changed its ad format.
- Image processing and OCR can be computationally expensive. You do not want to repeat these steps each time once you found some suitable solution. Thus, decide how to save your intermediate results afer this step, e.g., as external R, SQL, csv, or Excel files. You have to decide on the content and format of these files. Excel and csv files have the advantage to be human readable, so you can check their content more easily.

- Ultimately, answer the research questions you found suitable to be answered with your weekly ads. For ads in pdf format, it may be possible to do detailed price comparisons over an extended period of time. For ads in jpg, png, or another image format, it may not be possible to extract exact sales prices. Thus, such ads may only allow you to compare which products (and how many products) have been listed in a certain week.

- Think of efficient high–end visualizations of your final results. I would like to see at least two different meaningful high–end visualizations such as heat maps, timeline plots, scatterplot matrices, or parallel coordinate plots to mention only a few. The groups have been composed in such a way that each group contains at least two students who previously took (or concurrently take) *Statistical Visualization I*.

4 Timeline

- Each group has to give a short 4 to 5 min progress report each week, usually on Thursdays, specifically on 11/4, 11/11, and 11/18. In these short progress reports, you should present what you have done 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 11/4, each group has to describe which store ads they are going to use, how they plan to extract relevant information from these ads, and indicate which overall questions they expect to answer for their weekly ads.

  Two students are in charge of the progress report in a given week. Each student **must** be a presenter at least once for the weekly progress reports. The Zoom group can present via Zoom on the three dates listed above or send me a recording of their weekly progress report by 9am on these dates and I will play that recording in class.

- On Thursday 12/2, each group has to give a 20 min presentation of their final results via Zoom. This should contain a brief overview about the store chain you selected (e.g., headquarters, number of stores, number of employees, when founded, and any other interesting information you can find about the store). 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 graphical summaries. End with a discussion and outlook on things you would have liked to do, 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 planned to answer. Use my LaTeX Beamer slides (in Presentation.zip) as a template for this presentation.

  While there could be a "narrator" for this 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.

- On Friday 12/10 at 11:59pm, your six–page final written report is due. 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.

- On Saturday 12/11 at 11:59pm, your e–mail with your contributions and the percent–wise breakdown is due. See below for details.

5 Grading

The three progress reports in class are worth 9% of your total score, the final presentation in class on 12/2 is worth 40%, the final written report due on 12/10 is worth 50% of your total score for this project, and the e–mail with the breakdown of your contributions is worth 1%.

6 Additional Requirements

- It is up to you how to split the individual tasks. Each group member **must** present at least once 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 15% 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.

- Your 20 min presentation on Thursday 12/2 should be created via LaTeX Beamer. I will provide a template. You have to turn in your final resulting pdf file and the Rnw/tex file.

- A final six–page written report is due on Friday 12/10 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 12/11 at 11:59pm. You should not cc other group members on this e–mail and just send it to me. 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."
- **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 \bf Project Description} \\[1cm]
by \\[0.5cm]
{\bf J\"urgen Symanzik} \\[2.5cm]
{\bf Date:} \today \\[2cm]
{\bf Final Due Date:} Friday, December 10, 2021, 11:59pm (by e--mail) \\[2cm]

UTAH STATE UNIVERSITY \\[0.5cm]
Logan, UT \\[0.5cm]
Fall 2021 \\[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{General Instructions}

Your project is related to weekly advertisements from stores in the United States.
These could originate from local food stores / supermarkets in Utah
(such as Smith's, Macey's, Lee's Marketplace, etc.)
or from across the US (archived at \url{https://weekly-ads.us/}, dating back to 2018).
You could also work with ads from national general merchandise retailers
(such as WalMart, Target, ShopKo, etc.),
electronic store chains (such as Best Buy, etc.),
home improvement store chains (such as Lowe's, Home Depot, Menards, etc.),
or others.

The overall question is: How did weekly advertisements change
after the outbreak of the Coronavirus pandemic in March 2020
and the recent problems with the supply chain shortages in the
second half of 2021, i.e., how did these events affect what is
listed in weekly store advertisements?
For example, did toilet paper disappear from weekly ads in the
middle of 2020 and did bananas disappear from weekly ads in
the fall of 2021?

You have to work in groups of three to four students
on this project. Each group has to decide on its own local
or national chain. {\bf Once your group has decided on a chain,
claim your chain as quickly as possible to me via e--mail
and cc the other groups on your e--mail.} No other group
can claim the same chain thereafter.

There are three groups overall. Members of Groups 1 and 2 indicated that they
would prefer to meet with other group members in person. Members of Group 3
indicated that they would prefer to meet via Zoom. Ultimately, it is up to
each group to decide when (and how) to meet. Clearly, Groups 1 and 2 can
meet via Zoom as well, e.g., on weekends.

I consider course projects to be generally open to input from others.
All students are expected to frequently communicate and
closely work together on overlapping questions, even across groups
if that turns out to be helpful. Moreover, this project is open
to comments and suggestions from the other students in class as well.
So, you also have to consider suggestions made in class or in Canvas
by the other students from this course.
There are a few specific requirements and a timeline
you have to follow. More details in the sections below.


\section{Possible Research Questions}

The overall goal of weekly advertisements is to attract shoppers to a
store so that they purchase other non--sale items as well.
Since the start of the Coronavirus pandemic outbreak in March 2020,
our shopping behavior has changed considerably, e.g., far more
online shopping, home deliveries, or pickups on the outside of a store.
Moreover, there were supply shortages due to the pandemic,
```

```
and more recently, due to global supply chain shortages.
How did all of this affect what is advertised in weekly ads
(and at which price)?

Here is a list of possible questions
of interest to store managers and customers that are related to weekly ads.
This list easily could be extended.
You should select weekly ad data and then provide answers to some of these
questions. It is not expected that each group answers all possible questions.

\begin{itemize}
\item What should be placed on page 1 of a weekly ad?

\item What do competitors put on sale?

\item What are the prices of the advertised articles over time?

\item How many items are listed in an ad --- and how many on page 1? Did this change over time?

\item What is the median price across all items in the ad / on page 1? Did this change over time?

\item What are the different types of deals (33\% off, buy 6 --- save 60c on each,
buy 1 --- get one free or one half off, \$xx off, etc.)?

\item What type of information is provided to the customer, e.g., fixed price amount in
the ad vs.\ amount of savings in the ad?

\item What is the basis of the discount, e.g., on sale by unit (e.g., pound) vs.\ on sale by each?

\item Are there any extras, e.g., buy more, add points (e.g., for fuel, $\ldots$)?

\item What is the seasonality of advertised articles?

\item Global min / max for advertised items. When/where was this item cheapest / most expensive
over time across all locations?

\item How frequently are items on sale? Is there a certain pattern, e.g., every 4 or 5 weeks?

\item Are prices for advertised articles always the same, e.g., buy 3 for \$10 vs.\ \$3.99 each?

\item How did Covid affect ad prices (e.g., shortages, store closures / reduced hours, etc.)?

\item How did Covid affect which items are listed in ads?

\item Are there visible effects of supply chain shortages on items in ads during the past 2 or 3 months
(due to backlogs in US ports, truck driver shortages, etc.)?

\item Additional question you can think of or that are suggested by other students
from this course.
\end{itemize}


\newpage


\section{Specific Instructions}

These are the main components for this project:

\begin{itemize}
\item All your work most be done in R. You cannot manually manipulate files and data.
Your code should be reproducible/reusable, in particular for forthcoming weekly ads
that will be released while you are working on the project

\item Identify your chain. Then download their weekly ads in digital format.
```

```
Do this over an extended period of time with random time intervals in between, e.g., a few minutes.
Possibly do this from different computers, e.g., all 2021 ads from
computer 1, all 2020 ads from computer 2, etc. You do not want to
get blocked from further accesses because the ad server thinks
your ad downloads are a malicious attack.
Store each of the ads you downloaded locally on your harddrive,
or even better, in a cloud drive such as Box, Dropbox, or Google Drive
that can be shared among all group members.

\item Depending on the file format of the weekly ads, you may have
to use different approaches how to extract relevant product and price information.
For pdf files, it should be straight forward to extract relevant information
precisely and quickly. For jpg, png, or other image files, OCR will
be a major component.
Note that the {\tt WeeklyAdOutline.R} file discussed in class
should serve as a starting point and not as the end point to extract data
from image files. You have to see how you can best clean an ad page
from background noise (and images), make use of the bounding box
of a text, or see whether the reported confidence information for a text
could be used to filter out text that is not very reliable.

\item Regular expressions likely will play a major role to extract relevant data
and clean up the OCR results. But, do not forget to proofread a sample of
weekly ads. The group member who initially comes up with the regular expressions
is equally important as the group member who manually (i.e., visually)
compares the extracted OCR results with the actual ad pages and comes up with
shortfalls, omissions, and common interpretation problems of the OCR functionality.
Apparently, this information should be used to optimze and fine--tune the
regular expressions. Then, another visual assessment is necessary (and so on).
Once you feel comfortable that you get meaningful results for this subset
of ads, apply your methods to all ads. Even then, check for unexpected
results and outcomes, e.g., if the chain unexpectedly changed its ad format.

\item Image processing and OCR can be computationally expensive.
You do not want to repeat these steps each time once you found some
suitable solution. Thus, decide how to save your intermediate
results afer this step, e.g., as external R, SQL, csv, or Excel files.
You have to decide on the content and format of these files.
Excel and csv files have the advantage to be human readable, so you
can check their content more easily.

\item Ultimately, answer the research questions you found suitable
to be answered with your weekly ads. For ads in pdf format, it may be possible
to do detailed price comparisons over an extended period of time.
For ads in jpg, png, or another image format, it may not be possible to extract
exact sales prices. Thus, such ads may only allow you to compare which products (and how many products)
have been listed in a certain week.

\item Think of efficient high--end visualizations of your final results.
I would like to see at least two different meaningful high--end visualizations
such as heat maps, timeline plots, scatterplot matrices, or parallel coordinate plots
to mention only a few. The groups have been composed in such a way that each
group contains at least two students who previously took (or concurrently take)
{\it Statistical Visualization I}.

\end{itemize}


\section{Timeline}

\begin{itemize}
\item Each group has to give a short 4 to 5~min progress report each week,
usually on Thursdays, specifically on 11/4, 11/11, and 11/18.
In these short progress reports,
you should present what you have done 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 11/4, each group has to describe which store ads
they are going to use, how they plan to extract relevant information from these ads,
and indicate which overall questions they expect to answer for their weekly ads.

{\bf Two students} are in charge of the progress report in a given week.
Each student {\bf must} be a presenter at least once for the weekly progress reports.
The Zoom group can present via Zoom on the three dates listed above or
send me a recording of their weekly progress report by 9am on these dates
and I will play that recording in class.

\item On Thursday 12/2, each group has to give a 20~min presentation of their final results via Zoom.
This should contain a brief overview about the store chain you selected (e.g., headquarters, number of stores,
number of employees, when founded, and any other interesting information you can find about the store).
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 graphical summaries.
End with a discussion and outlook on things you would have liked to do,
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 planned to answer.
Use my \LaTeX\ Beamer slides (in Presentation.zip) as a template for this presentation.

While there could be a ``narrator'' for this 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 On Friday 12/10 at 11:59pm, your six--page final written report is due. 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.

\item On Saturday 12/11 at 11:59pm, your e--mail with your contributions and the
percent--wise breakdown is due. See below for details.

\end{itemize}


\section{Grading}

The three progress reports in class are worth 9\% of your total score,
the final presentation in class on 12/2 is worth 40\%, the final written report
due on 12/10 is worth 50\% of your total score for this project,
and the e--mail with the breakdown of your contributions is worth 1\%.


\section{Additional Requirements}

\begin{itemize}
\item It is up to you how to split the individual tasks.
Each group member {\bf must} present at least once 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 15\% 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 Your 20~min presentation on Thursday 12/2 should be created via
```

```
\LaTeX\ Beamer. I will provide a template. You have to turn in
your final resulting pdf file and the Rnw/tex file.

\item A final six--page written report is due on Friday 12/10 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 12/11 at 11:59pm.
You should not cc other group
members on this e--mail and just send it to me.
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.''

\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.