STAT 6080 Data Technologies

Project Description

by

Jürgen Symanzik

Date: November 15, 2020

Final Due Date: Wednesday, December 16, 2020, 11:59pm (by e–mail)

UTAH STATE UNIVERSITY

Logan, UT

Fall 2020

Contents

1 General Instructions

Your project is related to the Billboard Hot 100 charts, accessible at `https://www.billboard.com/charts/hot-100`. The charts legend at `https://www.billboard.com/p/billboard-charts-legend` provides background information how the Billboard charts are compiled. The Billboard Hot 100 charts were first released on August 4, 1958 (`https://www.billboard.com/charts/hot-100/1958-08-04`). A Wikipedia page (`https://en.wikipedia.org/wiki/Billboard_Hot_100`) contains interesting background information. As always with Wikipedia, do not fully rely on the information posted on their web pages, but always double check alternative sources. This web page may contain some information you might be able to partially replicate with your data: `https://www.insider.com/artists-most-number-one-songs-hot-100-2020-9`.

Questions one might want to answer are related to artists, songs, durations, and changing patterns, e.g., which artist had most songs on the charts overall, which artist spent most time at number 1, which artist had most songs on the charts in a single week, which songs spent most weeks on the charts, etc. What are the top–10 artists with respect to total weeks on the charts? Is there an artist who spent a single week on the charts on position 100? Numerous other questions could be answered.

You have to work in a group of two students on this project. All students are expected to frequently communicate and closely work together. However, I want to make this a project that is open to comments and suggestions from the other students in class as well. So, you have to consider suggestions in our Zoom discussions or in Canvas made by the other students from this course.

There are a few specific requirements and a timeline you have to follow. More details follow in the sections below.

2 Possible Research Questions

About 62 years $\times$ about 52 weekly charts, i.e., more than 3,000 Hot 100 charts have been released so far. This results in more than 300,000 charts entries. You have to mine all of them! You can start at the first or very last released version of the charts and then move forward or backward through the released charts. You can also access the charts in batches of 5 or 10 years. Keep track of which charts pages you have scraped already. In case your webscraping gets blocked by the Billboard web server, try again in smaller batches and with wider time intervals between consecutive accesses to the server.

Here are some possible questions to investigate:

- Which Artist was featured on the charts the most, i.e., which artist occurs the most? And what are the top–10 or top–50 artists with most weeks on the charts?
- Which song spent the most time on the charts?

- Which song spent the most time in the number 1 position?
- Which artist had music on the charts over the longest amount of time?
- Basic questions such as "What were the top 10 on a given date".
- Has the average amount of time a song has spent on the charts changed over time?
- Which songs have charted by a certain artist? Restrict this list to the single artist or include collaborations as well.
- What are the most frequent words in song titles? Love?!? Consider to remove some stop words such as "a" and "the" using the *stopwords* R package, but do not remove all of them.
- Similar questions, but now restricted to a certain year or a certain time period, e.g., "in 2019" or "in the 1990ies".
- Additional questions you can think of or that are suggested by other students from this course.

Be careful with artist names and song titles. An artist name may be a substring of another artist name or part of a collaboration. Here is an example for the substring "cher":

- Cher: `https://en.wikipedia.org/wiki/Cher`
- Sonny & Cher: `https://en.wikipedia.org/wiki/Sonny_%26_Cher`
- Cher and Peter Cetera: `https://en.wikipedia.org/wiki/After_All_(Cher_and_Peter_Cetera_song)`
- Cherie (Cyndi Almouzni): `https://en.wikipedia.org/wiki/Cyndi_Almouzni`
- Cherish: `https://en.wikipedia.org/wiki/Cherish_(group)`
- Buckcherry: `https://en.wikipedia.org/wiki/Buckcherry`

There are three main components to this project:

- Webscrape and process the about 3,000 Billboard Hot 100 web pages. Extract relevant information using regular expressions if necessary. You can use all R packages of your choice, even if not discussed in class. You should download the web pages once and then no further access the Billboard web server.

- Store the relevant information on your local computer or in box or dropbox. It is up to you which format to choose, e.g., about 3,000 single files with 100 rows (one for each week), files for a year or multiple years, or a single file with about 300,000 rows. The file format is up to you, e.g., an Excel csv or xlsx format or an SQL data base format.

  Your file format should match how you are going to extract the data from your webscraping attempts and how you are going to load the data into R, e.g., via SQL queries or by loading the entire data as a data frame or tibble into R. In the latter case, consider carefully how much memory will be used and whether it might be better to store artist names and song titles as factors or as strings.

- Develop a simple user interface that allows a user to type in questions such as "*How many weeks did Cher spend on the charts?*" or "*List all song titles by Cher that appeared on the charts*" or "*What is the most frequent word in all song titles? Ignore stop words!*". Think of other likely questions and ask the class for suggestions. Using regular expressions, translate these human questions into SQL queries or tidyverse functionality that can be used to extract the relevant information from your data.

  Look at the *question* function from the 2019 Marvel project and use this as a starting point for your own solution to answer user queries related to the Billboard Hot 100 charts. R isn't fully capable of natural language support, but first steps have been made, see `https://yihui.shinyapps.io/voice/`. It may not be too long that we communicate with an Rlexa interface that translates some of our everyday language into R expressions.

3 Specific Instructions

- All your work most be done in R. You cannot manually manipulate files and data. Your code should be reproducible/reusable, in particular as additional web pages are created on a weekly basis. Rather access all pages that exist when you run the final version of your R code. Also assume that someone may run your code again in a year with new artist names, new collaborations, and new song titles.

- Download all web pages just once. Decide in advance how to process your downloaded web pages and in which format to store the information.

- Create a simple user interface where a human can enter a specific question. This question gets translated via regular expressions into an SQL query or tidyverse functionality to extract the relevant information from your data. You should allow questions that are case insensitive and do not depend on small spelling differences (e.g., a space or hyphen between two–word names and using "&" or "and" or "with"). Details are left to the group and can be decided during the progression of the project.

- Start exploring and processing some of the web pages. Initially work with a small sample of 20 or 30 web pages. Try different web pages across the entire 60 years and not only the most recent ones or the earliest ones. Then refine and adjust your process and check your results for the next 20 or 30 web pages. Adjust again if necessary. Finally process all of the about 3,000 web pages. This may be an overnight process without any further human interaction.

- Demonstrate that the user questions (as outlined above) will be answered in a meaningful way. Verify some of the results from the web pages mentioned earlier on. Other students (and professors) should be allowed to ask questions as well. So, if someone does not use a question mark at the end of a question, this still should be recognized as a valid question. All meaningful questions should result in an answer (possibly "I do not understand this question.").

4 Timeline

- You have to give a short 10 to 15 min progress report each week, usually on Thursdays until 11/19. In these short progress reports, you should summarize 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.

- On Tuesday 12/1, you have to give a 30 min presentation of your final results. This should contain a brief introduction into the Billboard Hot 100 charts (do a Google scholar search and note that there exist several scientific articles that deal with this topic, one even in JASA! — see `https://www.tandfonline.com/doi/abs/10.1198/016214501753168091`). The main part should be a summary of the computational methods and R packages you have used, and obviously a presentation of your results. 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. End with a live demonstration where other students and I can type our questions into the Zoom Chat window and you copy and paste these questions into your R "question" function. It is not necessary to come up with final answers to all initial questions. 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 Wednesday 12/16 at 11:59pm, your final written report is due. Likely you have to make a careful decision what to present 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.

5 Grading

The final presentation in class on 12/1 is worth 40% and the final written report due on 12/16 is worth 60% of your total score for this project.

6 Additional Requirements

- It is up to you how to split the individual tasks.
- 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 60% of the overall work for this project. A group member that contributes less than 40% of the overall work to this project will get individual point deductions. You have to provide me (via e–mail) with an estimated percentage of everyone's contributions to this project and the kind of contributions everyone made.
- Your 30 min presentation on Tuesday 12/1 should be created via LaTeX Beamer. Use my LaTeX Beamer slides (in Presentation.zip) as a template for this presentation. You have to turn in your final resulting pdf file and the Rnw/tex file.
- A final 6–page written report is due on Wednesday 12/16 at 11:59pm. This has to follow the JSM formatting requirements from the American Statistical Association (ASA). These 6 pages must contain everything from an abstract, keywords, the main sections of your report, all figures and tables, and the references. Use my LaTeX template (in ProceedingsPaper.zip) as a template for this final report. 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 immediately after the submission of your written report. You should not cc other group members in 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, reach out to me via e–mail or schedule a Zoom meeting. 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}

%\SweaveOpts{concordance=TRUE}

\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:} Wednesday, December 16, 2020, 11:59pm (by e--mail) \\[2cm]


UTAH STATE UNIVERSITY \\[0.5cm]
Logan, UT \\[0.5cm]
Fall 2020 \\[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  the
Billboard Hot 100 charts, accessible at
\url{https://www.billboard.com/charts/hot-100}.
The charts legend at
\url{https://www.billboard.com/p/billboard-charts-legend}
provides background information how the Billboard charts
are compiled.
The Billboard Hot 100 charts were first released
on August 4, 1958
(\url{https://www.billboard.com/charts/hot-100/1958-08-04}).
A Wikipedia page
(\url{https://en.wikipedia.org/wiki/Billboard_Hot_100})
contains interesting background information.
As always with Wikipedia, do not fully rely on the information posted
on their web pages, but always double check alternative sources.
This web page may contain some information you might be
able to partially replicate with your data:
\url{https://www.insider.com/artists-most-number-one-songs-hot-100-2020-9}.

Questions one might want to answer are related to artists,
songs, durations, and changing patterns, e.g.,
which artist had most songs on the charts overall, which artist
spent most time at number 1, which artist had most songs
on the charts in a single week, which songs spent most weeks on the charts, etc.
What are the top--10 artists with respect to total weeks on the charts?
Is there an artist who spent a single week on the charts on position 100?
Numerous other questions could be answered.

You have to work in a group of two students
on this project.
All students are expected to frequently communicate and
closely work together.
However, I want to make this a project that is open
to comments and suggestions from the other students in class as well.
So, you have to consider suggestions in our Zoom  discussions or in Canvas made
by the other students from this course.

There are a few specific requirements and a timeline
you have to follow. More details follow in the sections below.


\section{Possible Research Questions}

About  62 years $\times$ about 52 weekly charts, i.e., more than 3,000
Hot 100 charts have been released so far.
This results in more than 300,000 charts entries.
You have to mine all of them! You can start at the
first or very last released version of the charts and then move
```

```
forward or backward through the released charts.
You can also access the charts in batches of  5 or 10 years.
Keep track of which charts pages you have scraped already.
In case your webscraping gets blocked by the Billboard web server, try again
in smaller batches and with wider time intervals between consecutive accesses
to the server.

Here are some possible questions to investigate:
\begin{itemize}
\item Which Artist was featured on the charts the most, i.e., which artist occurs the most?
And what are the top--10 or top--50 artists with most weeks on the charts?
\item Which song spent the most time on the charts?
\item Which song spent the most time in the number 1 position?
\item Which artist had music on the charts over the longest amount of time?
\item Basic questions such as ``What were the top 10 on a given date''.
\item Has the average amount of time a song has spent on the charts changed over time?
\item Which songs have charted by a certain artist? Restrict this list to the single
artist or include collaborations as well.
\item What are the most frequent words in song titles? Love?!? Consider to remove some
stop words such as ``a'' and ``the'' using the {\it stopwords} R package,
but do not remove all of them.
\item Similar questions, but now restricted to a certain year or a certain
time period, e.g., ``in 2019'' or ``in the 1990ies''.
\item Additional questions you can think of or that are suggested by other students
from this course.
\end{itemize}

Be careful with artist names and song titles.
An artist name may be a substring of another artist name or part of a collaboration.
Here is an example for the substring ``cher'':
\begin{itemize}
\item Cher: \url{https://en.wikipedia.org/wiki/Cher}
\item Sonny \& Cher: \url{https://en.wikipedia.org/wiki/Sonny_%26_Cher}
\item Cher and Peter Cetera: \url{https://en.wikipedia.org/wiki/After_All_(Cher_and_Peter_Cetera_song)}
\item Cherie (Cyndi Almouzni): \url{https://en.wikipedia.org/wiki/Cyndi_Almouzni}
\item Cherish: \url{https://en.wikipedia.org/wiki/Cherish_(group)}
\item Buckcherry: \url{https://en.wikipedia.org/wiki/Buckcherry}
\end{itemize}

There are three main components to this project:
\begin{itemize}
\item Webscrape and process the about 3,000 Billboard Hot 100 web pages.
Extract relevant information
using regular expressions if necessary.
You can use all R packages of your choice, even if not
discussed in class. You should download the web pages once and then no
further access the Billboard web server.

\item Store the relevant information on your local computer or in box or dropbox.
It is up to you which format to choose, e.g., about 3,000 single
files with 100 rows (one for each week), files for a year or multiple years,
or a single file with about 300,000 rows. The file format
is up to you, e.g., an Excel csv or xlsx format or an SQL data base format.

Your file format should match how you are going to extract the data from
your webscraping attempts and how you are going to
load the data into R, e.g., via SQL queries
or by loading the entire data as a data frame or tibble into R.
In the latter case, consider carefully how much memory will be used and
whether it might be better to store artist names and song titles
as factors or as strings.

\item Develop a simple user interface that allows a user to type in questions such as
``{\it How many weeks did Cher spend on the charts?}'' or
``{\it List all song titles by Cher that appeared on the charts}'' or
```

```
``{\it What is the most frequent word in all song titles? Ignore  stop words!}''.
Think of other likely questions
and ask the class for suggestions. Using regular expressions, translate these
human questions into SQL queries or tidyverse functionality
that can be used to extract the relevant information from your data.

Look at the {\it question} function from the 2019 Marvel project and use this
as a starting point for your own solution to answer user queries
related to the Billboard Hot 100 charts.
R isn't fully capable of natural language support, but first steps have been
made, see \url{https://yihui.shinyapps.io/voice/}. It may not be too long
that we communicate with an Rlexa interface that translates some
of our everyday language into R expressions.
\end{itemize}


\section{Specific Instructions}

\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 as additional web pages
are created on a weekly basis. Rather access all pages that exist when
you run the final version of your R code. Also assume that someone may run
your code again in a year with new artist names, new collaborations,
and new song titles.

\item Download all web pages just once.
Decide in advance how to process your downloaded web pages
and in which format to store the information.

\item Create a simple user interface where a human can enter a specific question.
This question gets translated via regular expressions into an SQL query
or tidyverse functionality
to extract the relevant information
from your data. You should allow questions that are case insensitive and do not depend
on small spelling differences (e.g., a space or hyphen between two--word names
and using ``\&'' or ``and'' or ``with'').
Details are left to the group and can be decided during the progression of the project.

\item Start exploring and processing some of the web pages.
Initially work with a small sample of 20 or 30 web pages.
Try different web pages across the entire 60 years and not only
the most recent ones or the earliest ones.
Then refine and adjust your process and check your results for the next 20 or 30
web pages. Adjust again if necessary.
Finally process all of the about 3,000 web pages.
This may be an overnight process without any further human interaction.

\item Demonstrate that the user questions (as outlined above) will be
answered in a meaningful way.
Verify some of the results from the web pages mentioned earlier on.
Other students (and professors) should be
allowed to ask questions as well. So, if someone does not
use a question mark at the end of a question, this still should be
recognized as a valid question. All meaningful questions
should result in an answer (possibly ``I do not understand this question.'').

\end{itemize}


\section{Timeline}

\begin{itemize}
\item You have to give a short 10 to 15~min progress report each week,
usually on Thursdays until 11/19. In these short progress reports,
```

```
you should summarize 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.

\item On Tuesday 12/1, you have to give a 30~min presentation of your final results.
This should contain a brief introduction into the Billboard Hot 100 charts
(do a Google scholar search and
note that there exist several scientific articles that deal with this topic,
one even in JASA! ---
see \url{https://www.tandfonline.com/doi/abs/10.1198/016214501753168091}).
The main part should be a summary of the
computational methods and R packages you have used, and obviously a presentation of your results.
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.
End with a live demonstration where other students and I can type our questions
into the Zoom Chat window and you copy and paste these questions into
your R ``question'' function.
It is not necessary to come up with final answers to all initial questions.
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 Wednesday 12/16 at 11:59pm, your final written report is due. Likely you have
to make a careful decision what to present 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.

\end{itemize}


\section{Grading}

The final presentation in class on 12/1 is worth 40\% and the final written report
due on 12/16 is worth 60\% of your total score for this project.


\section{Additional Requirements}

\begin{itemize}
\item It is up to you how to split the individual tasks.

\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 60\% of the overall work for this project. A group
member that contributes less than 40\% of the overall work to this project
will get individual point deductions. You have to provide me (via e--mail) with
an estimated percentage of everyone's contributions to this project
and the kind of contributions everyone made.

\item Your 30~min presentation on Tuesday 12/1 should be created via
\LaTeX\ Beamer.
Use my \LaTeX\ Beamer slides (in Presentation.zip) as a template for this presentation.
You have to turn in
your final resulting pdf file and the Rnw/tex file.

\item A final 6--page written report is due on Wednesday 12/16 at 11:59pm. This has
to follow the JSM formatting requirements from the American Statistical Association (ASA).
These 6 pages must contain everything from an abstract, keywords, the main
sections of your report, all figures and tables, and the references.
Use my \LaTeX\ template (in ProceedingsPaper.zip) as a template for this final report.
```

```
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 immediately after the
submission of your written report. You should not cc other group
members in 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, reach out to me
via e--mail or schedule a Zoom meeting. 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.