Data Visualization

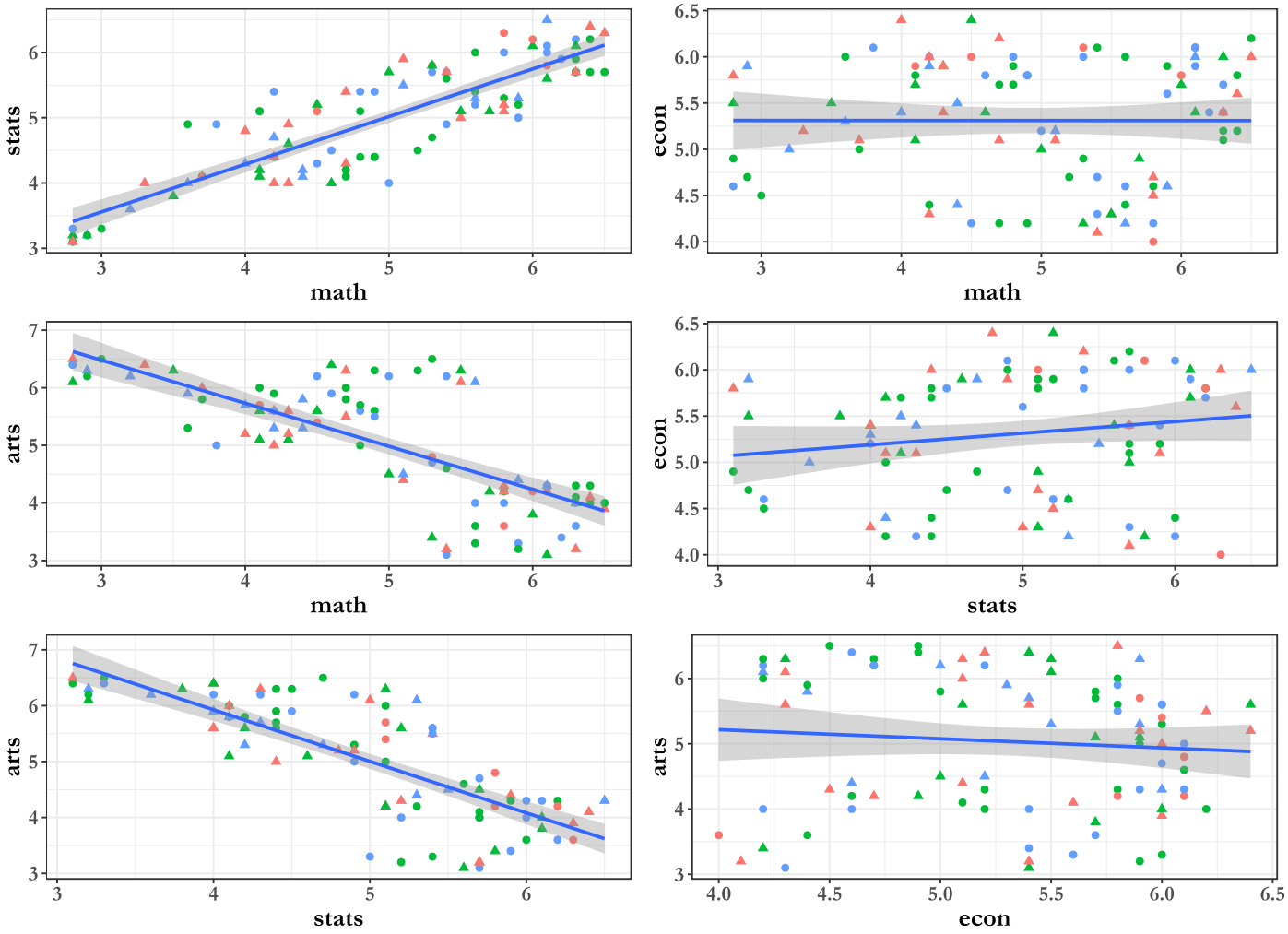
Portfolio 2019

Noah Mamié
Splügenstrasse 7
9008 St. Gallen
17-607-714

4,784 Skills: Data Visualization | Peter Gruber | University of St. Gallen

Contents

Student Data



This arrangement of correlations presents a conslusion of this year's student grades derived from various courses.

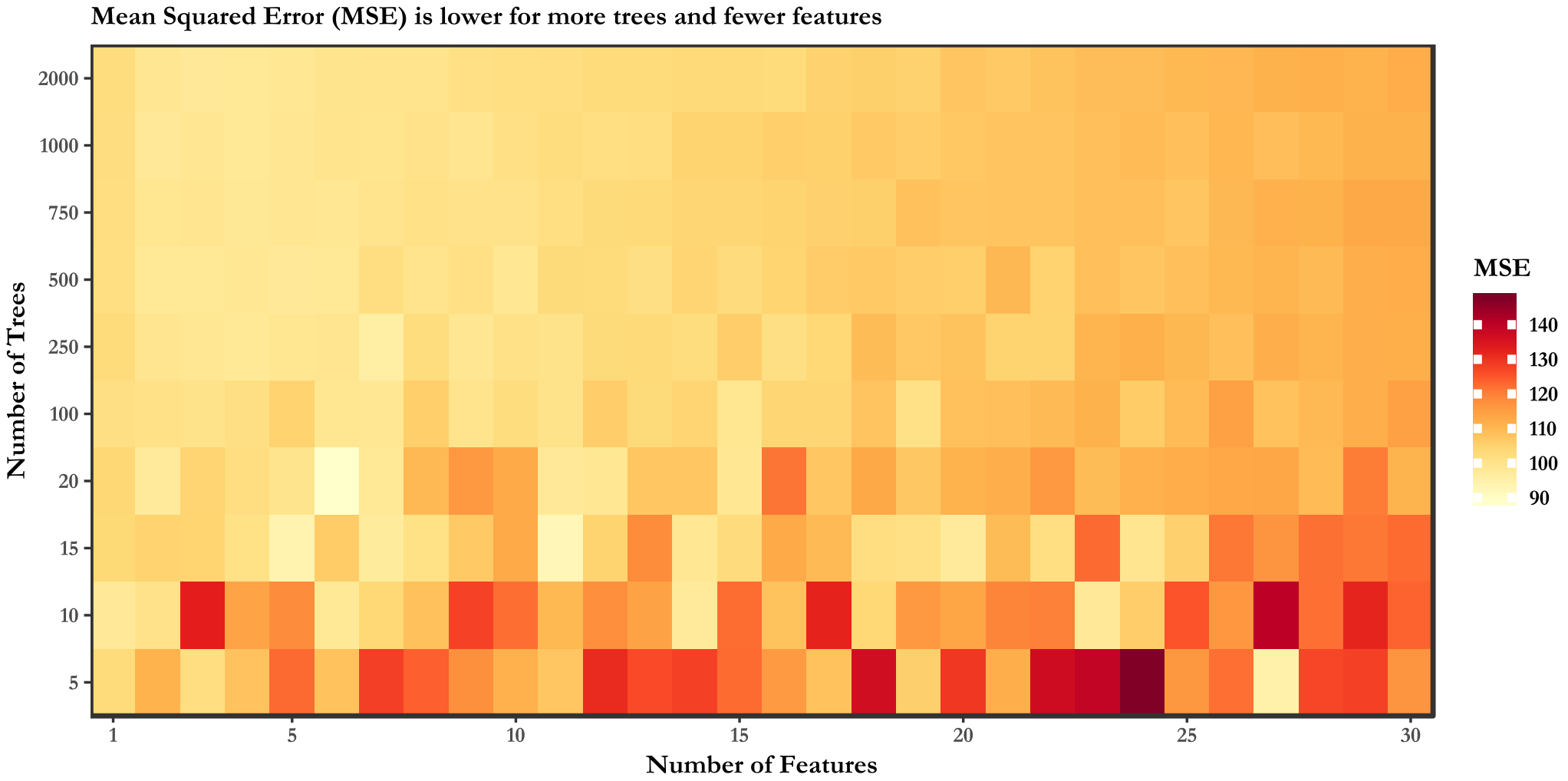
Country

- Austria
- Switzerland
- Germany

Gender

- Female
- Male

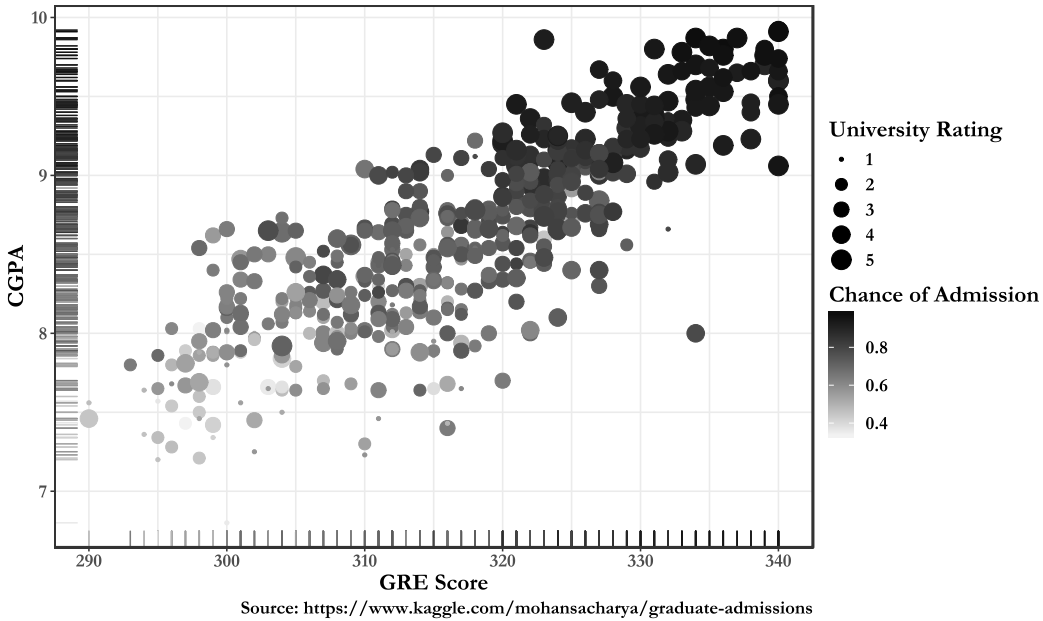
Random Forest Performance on Testing Data

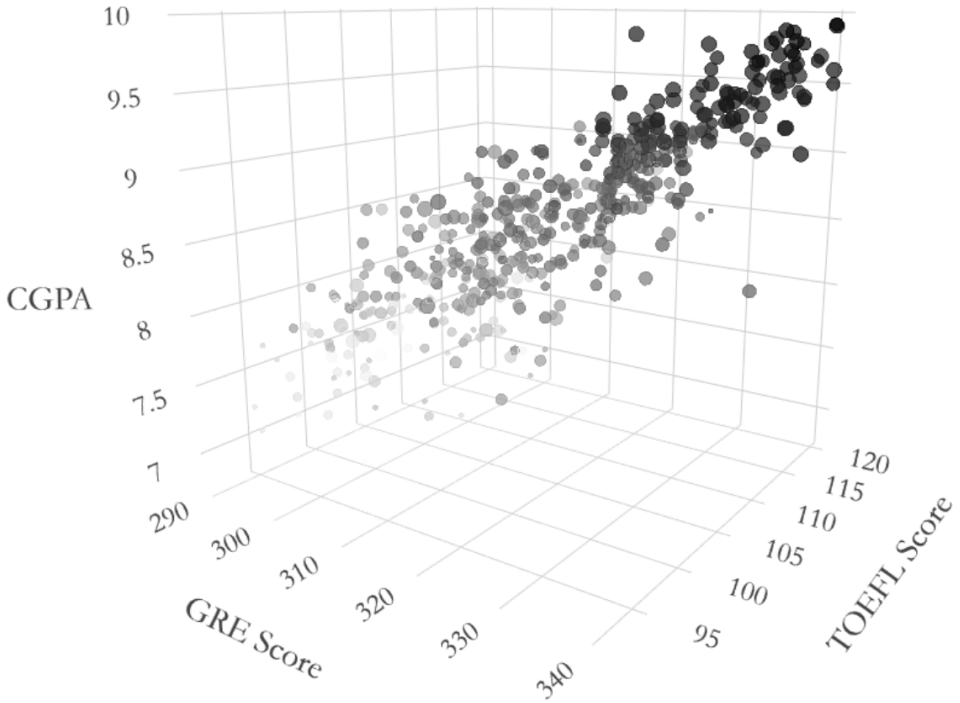


Source: https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

Prediction of Graduate Admission

"The most colorful thing in the world is black and white, it contains all colors and at the same time excludes all."
~Vikram Verma (2013)





This 3D plot for the 'Graduate Admission' dataset can be further investigated here:
https://plot.ly/~TheTrueHOOHA/23/

Twitter Statistics

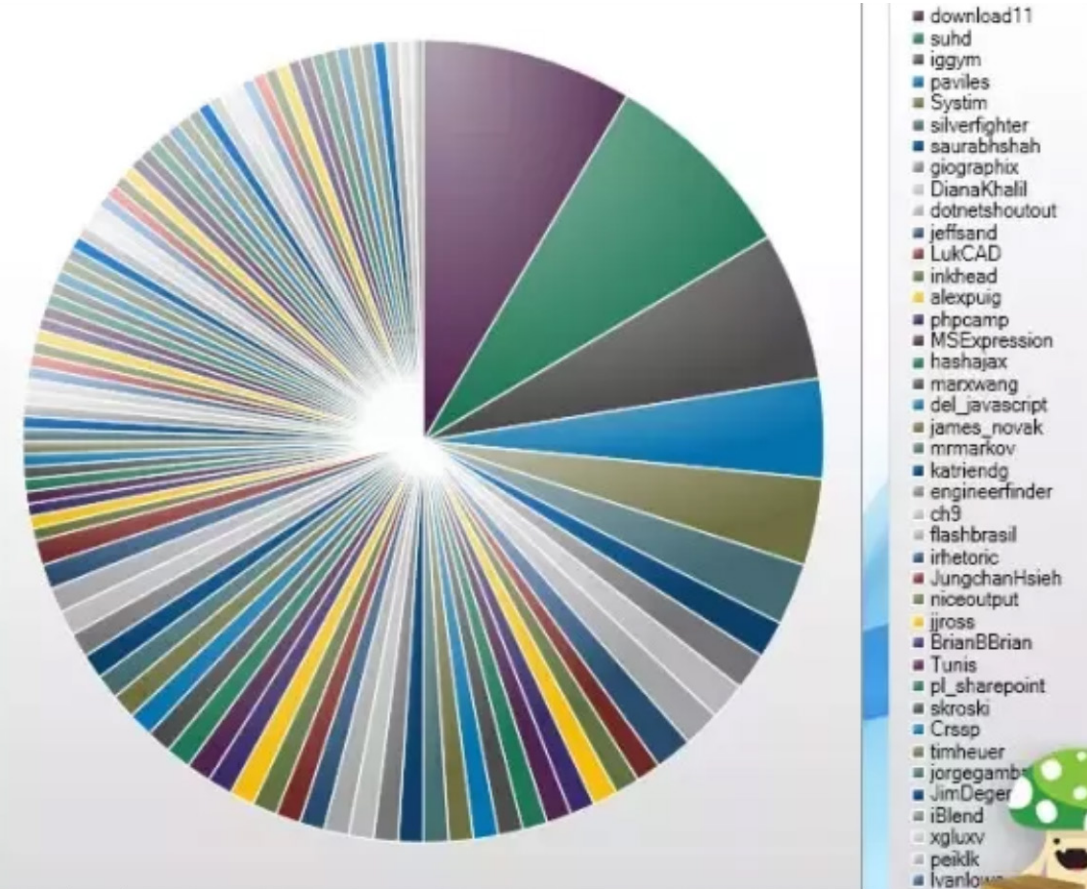
This pie chart is probably one of the worst attempts to meaningfully visualize a set of data. The intention of displaying the distribution of tweets between the most active Twitter users was in no way achieved, as it is impossible to make sense of this plot.

First of all, the dimension of the tweets sent by the users is not recognizable at all. Further, there are no numbers displayed on the chart and also the names are only visible partly as a list next to the chart.

From a content perspective, the chart is simply to packed with data to be readable. Thus, it is more than just questionable whether the use of a pie chart for this visualization was justified.

Finally, the aesthetics and the quality of the plot are not appealing. This impression is probably triggered by the strange shiny center and the low resolution.

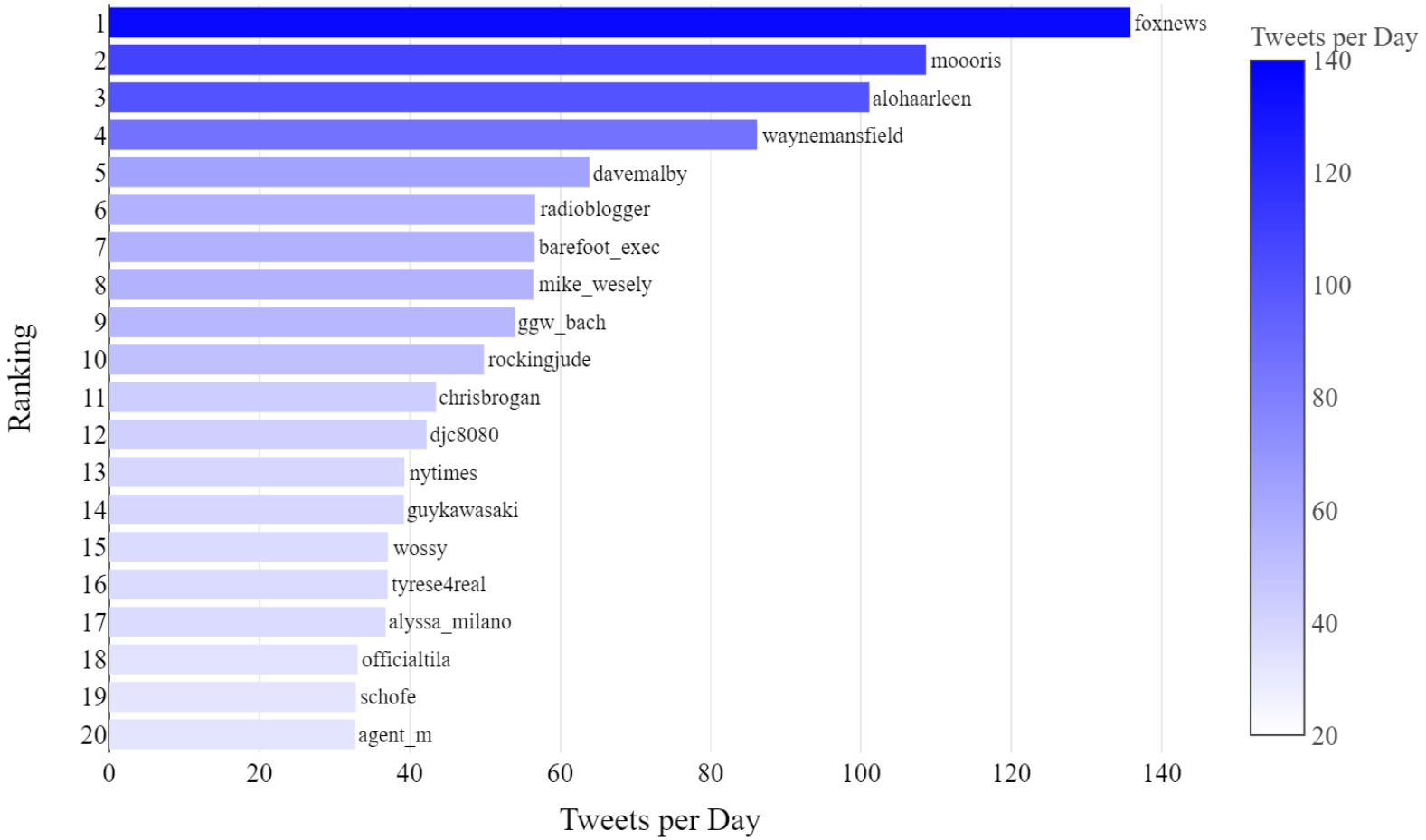
Most Active Twitter Users



Source: https://chandoo.org/wp/nightmarish-pie-charts/

Twitter Statistics

Top 20 Most Active Twitter Users in 2019



Visualizing a ranking dataset with a bar plot proves to be the better solution than a pie chart. The readability is very high and, additionally, the high density of information does not present an issue anymore.

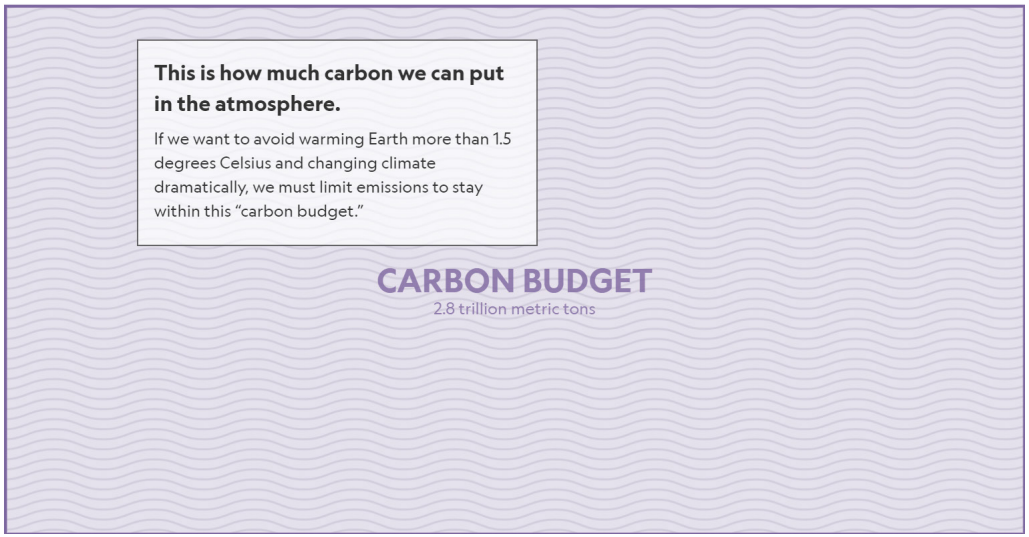
This advantage of a barplot is even more impressing as an interactive shiny app:
https://dataviz19-noah.shinyapps.io/Twitter_Most_Active/

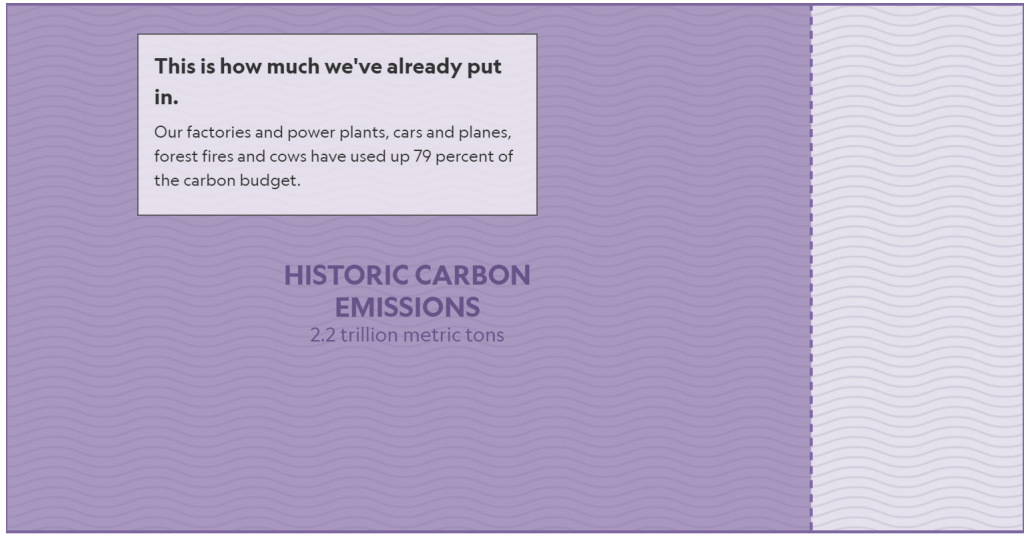
Carbon Budget - Storytelling

A good visualization should be able to tell a story in a meaningful way, enabling a better understanding of the data presented.

On nationalgeographic.com under the section climate engineering there are many graphics that intend to raise awareness about the climate change. Especially the simplicity of this particular chart about our planet's carbon budget struck me. Although the consecutively introduced charts do not possess a high information density, the story they communicate is still extremely relevant and impressive.

Firstly, the creator of this storyline present current information about the carbon budget and the respective danger is poses to our environment.





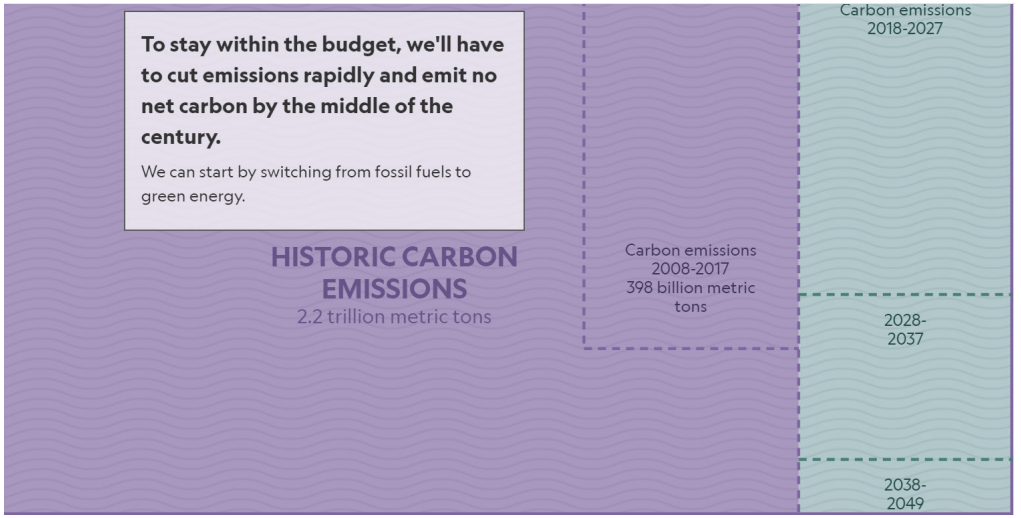
Carbon Budget - Storytelling

After having presented further information about historic carbon emissions the last step involves a conclusion and a calculation that leads to a solution for future emissions.

The result, while being a bit shocking, is thoroughly evaluated in the solution approach and thereby sparks a tiny bit of hope.

This interactive visualization can be found here: https://www.nationalgeographic.com/environment/2018/12/climate-geoengineering-series-intro/

NOTE: Projections are based on the median values of 42 scenarios with low or limited overshoot of the remaining carbon budget at a 50 percent probability.



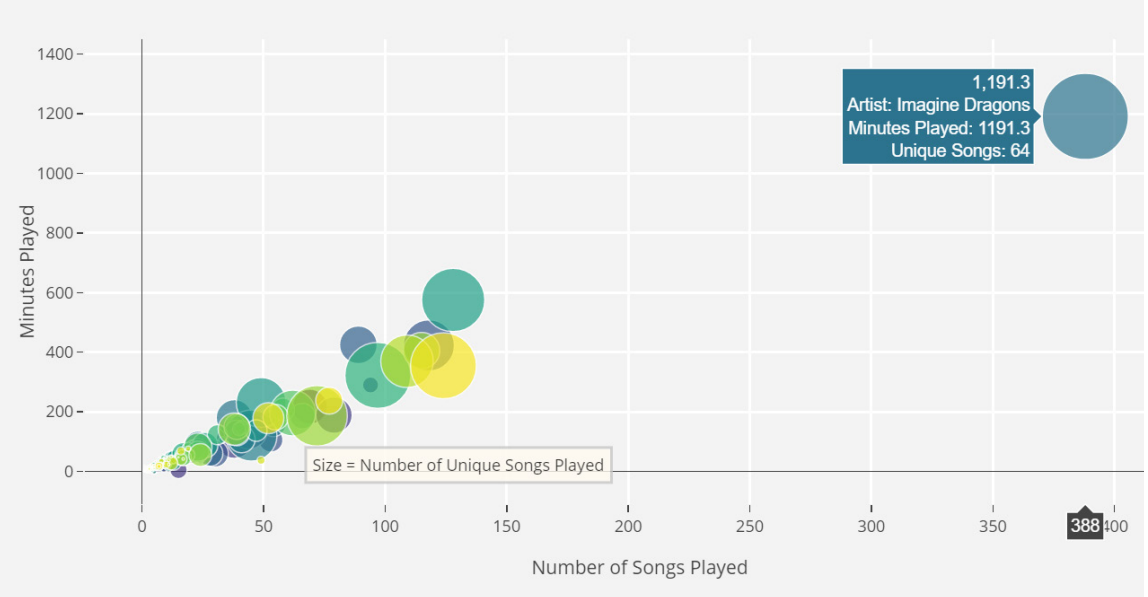
Spotify Data

Listening to all different kinds of music is one of my most important daily routine. Therefore, analyzing my personal *Spotify* data presented a special opportunity for me, getting to know the details of my listening habits.

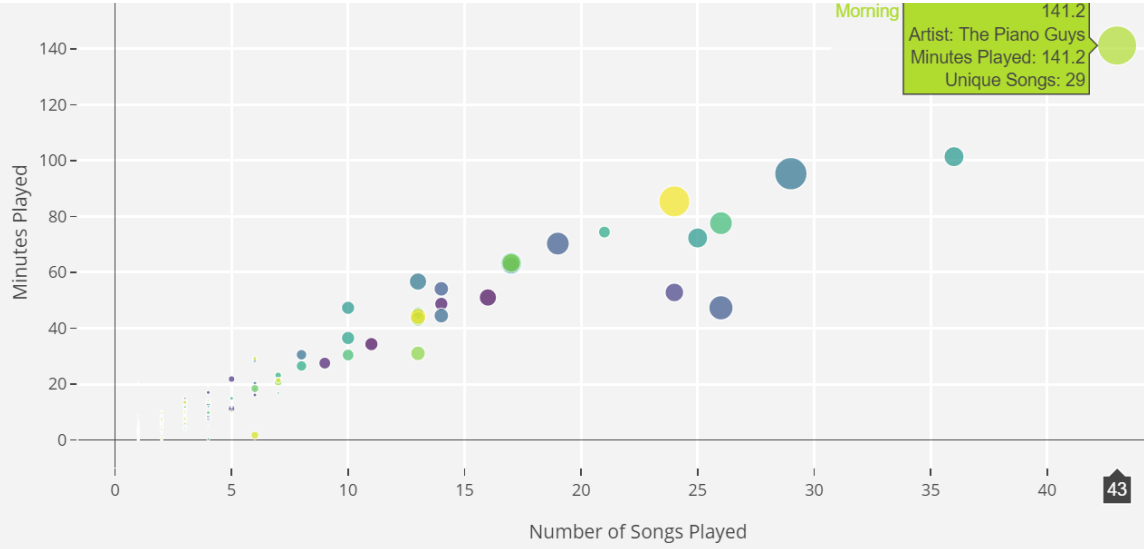
Looking at the bubble chart plot it is depicted that overally I have a tendency to listen to *Imagine Dragons* quite a lot (cf. upper picture). However, filtering the data for the different daytimes it is noticeable that my musical preferences vary across the hours of the day. Apparently, in the morning I prefer listening to instrumental music, e.g. *The Piano Guys*, while in the afternoon/evening vocal music takes over my preferences.

To have a more detailed insight into the my listening habits on *Spotify* please follow the link to the interactive bubble chart created with *Plotly*:
https://plot.ly/~TheTrueHOOHA/14/artist-variance-by-time-of-day/

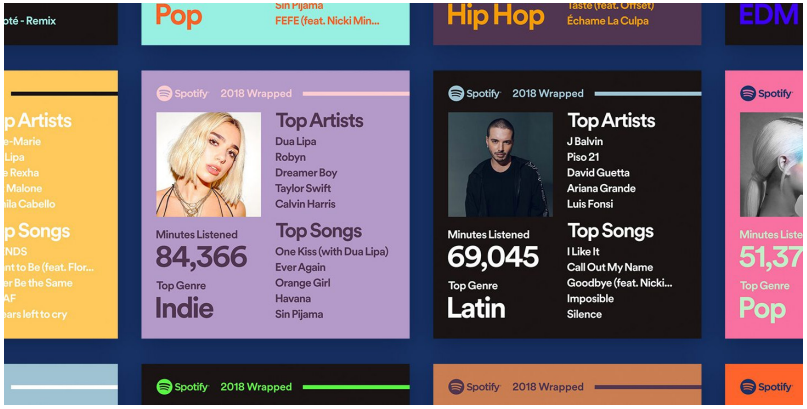
Amount of Time Listened and Songs by Artist



Morning



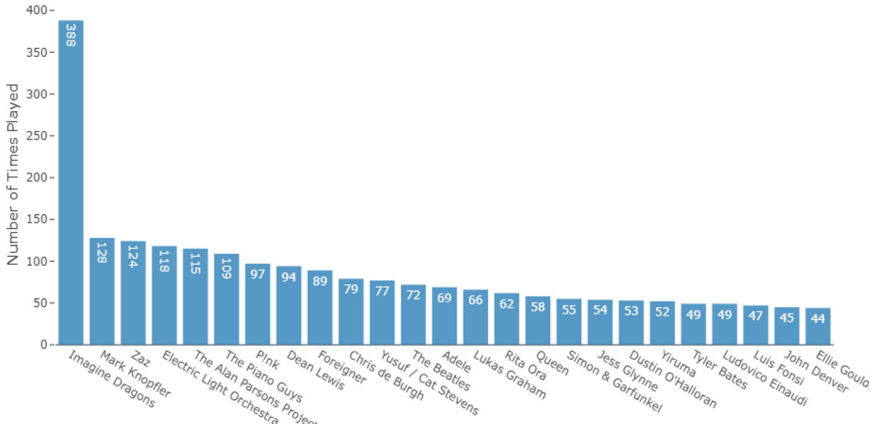
Creation Process



The inspiration for this visualization came from Spotify's feature "Spotify Wrapped". This tool provides users with their personal data, i.e. favorite songs, minutes played etc.

1

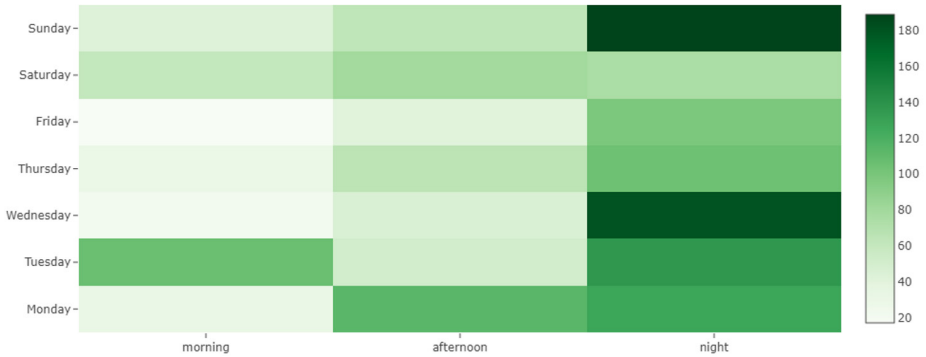
Popularity of Artist by Count



First of all, I started my approach with a vertical bar chart that displays my personal preferences when choosing artists to listen to. The data for this chart (and the ones to follow) was downloaded from my personal Spotify account and then processed in a Python Jupyter Notebook.

```
data = [
  go.Bar(
    x=most_popular_artists_by_count.index,
    y=most_popular_artists_by_count,
    text=most_popular_artists_by_count,
    textposition='auto',
    opacity=0.75)]
layout = go.Layout(
  title='Popularity of Artists by Count',
  yaxis= dict(
    title='Number of Times Played',
    gridcolor='rgb(255, 255, 255)',
    zerolinewidth=1,
    ticklen=5,
    gridwidth=2,
    titlefont=dict(size=15)))
fig = go.Figure(data=data, layout=layout)
py.iplot(fig, filename='popular_artists')
```

2

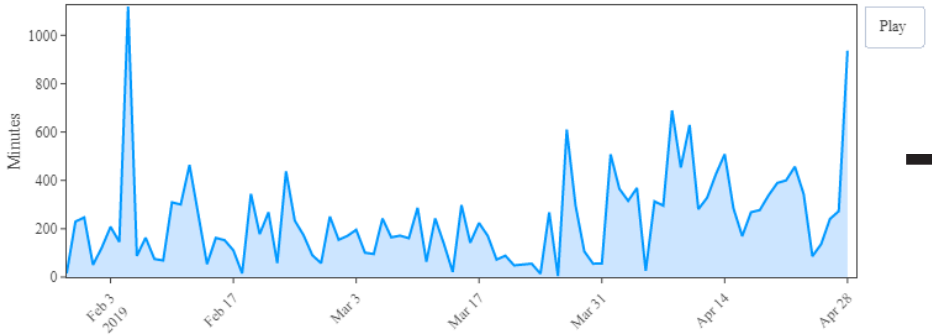


The second step involved the creation of a heatmap that illustrates my personal spotify usage based on the weekday and the time. This proved that I primarily listen to music in the evening.

```
trace = go.Heatmap(z=time_of_day_utc_pivot.values,
                   x=time_of_day_utc_pivot.columns,
                   y=time_of_day_utc_pivot.index,
                   colorscale='Greens',
                   reversescale=True)
data=[trace]
py.iplot(data, filename='spotify_heatmap')
```

Creation Process

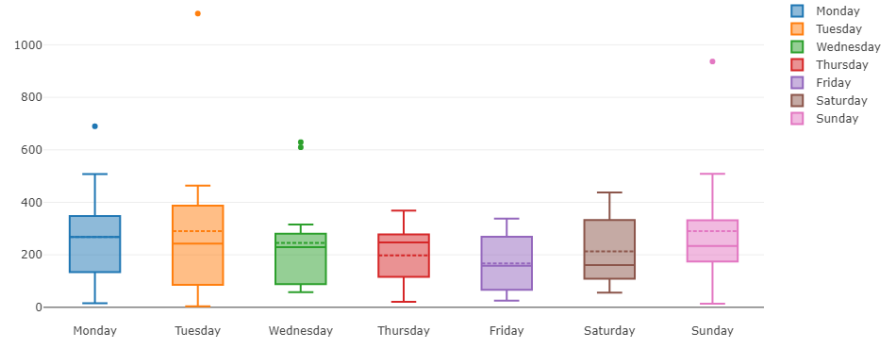
3 Minutes Played per Day Spotify



Thirdly, the analysis involved the total minutes played per day over the last three months, showing values of more than 1000 minutes per day.

```
layout = dict(title='Minutes Played per Day Spotify',
              font=dict(family='Garabd'),
              showlegend=False,
              autosize=False,
              width=800,
              height=400,
              xaxis=dict(axis, **{'nticks':12, 'tickangle':-45,
                'range': [min(number_of_minutes_per_day.index),
                          max(number_of_minutes_per_day.index)]}),
              yaxis=dict(axis, **{'title': 'Minutes',
                'range':[0,max(number_of_minutes_per_day.minutes_played)+5]})
```

4 Minutes Played per Weekday Boxplot



The fourth and penultimate step of the process was the creation of boxplots for the individual weekdays that illustrate the minutes played per weekday.

```
days = ['Monday','Tuesday','Wednesday','Thursday','Friday','Saturday','Sunday']

time_per_day = []

for day in days:
  df_days =
  merged_minutes_per_day[merged_minutes_per_day.local_day_of_week == day]
time_per_day.append(list(df_days['minutes_played']))

traces = []

for i, j in enumerate(days):
  trace = go.Box(
    y=time_per_day[i],
    name = j,
    boxmean=True)

traces.append(trace)
data = traces
py.iplot(data)
```

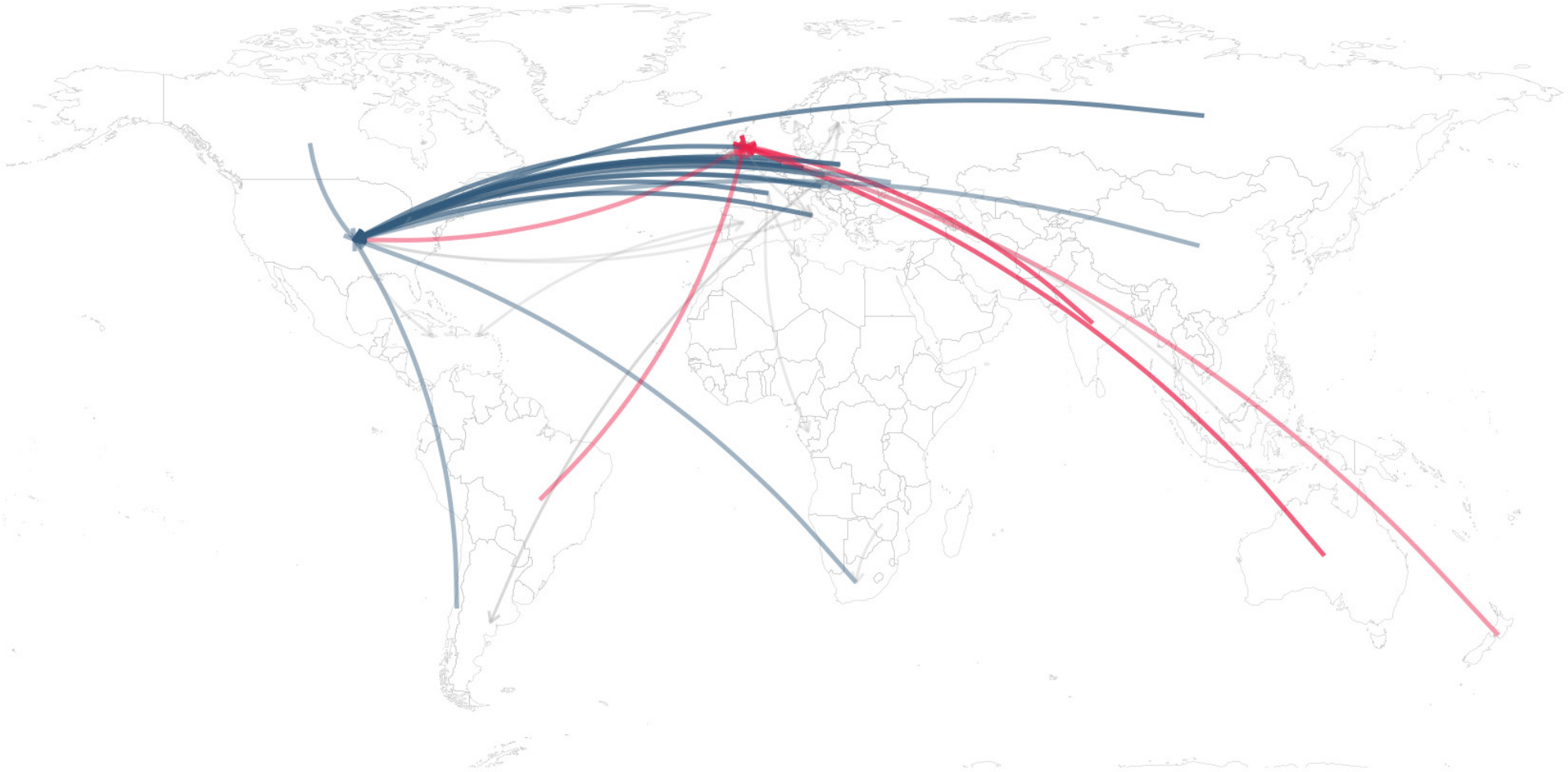
5

Finally, the thereby created plots were assembled and, in its entirety, present my personal Spotify usage in the form of a bubble chart. Additionally, the artist variance by time of day may be observed.

```
scatter_morning = time_of_day_scatter_df(time_of_day='morning')
scatter_afternoon = time_of_day_scatter_df(time_of_day='afternoon')
scatter_evening = time_of_day_scatter_df(time_of_day='night')
```

Nobel Prize Winners

73% of Those Dying Abroad Pass Away in 1 of 5 Countries

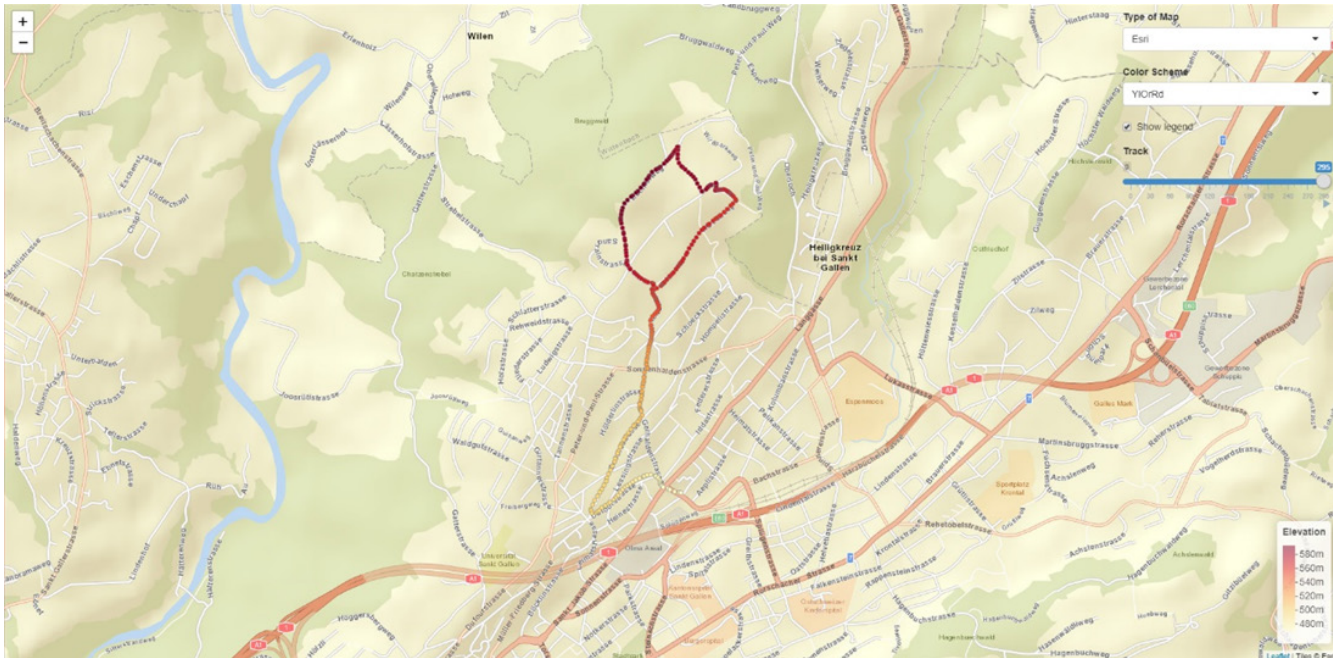


The interactive visualization created with shinyapps.io can be found under the following link:
https://themachine.shinyapps.io/nobel/

Garmin Connect Data Map

These maps illustrate my personal jogging route in St. Gallen. The data points have been generated with a Garmin Vivoactive HR and the dataset was provided from the official Garmin website.

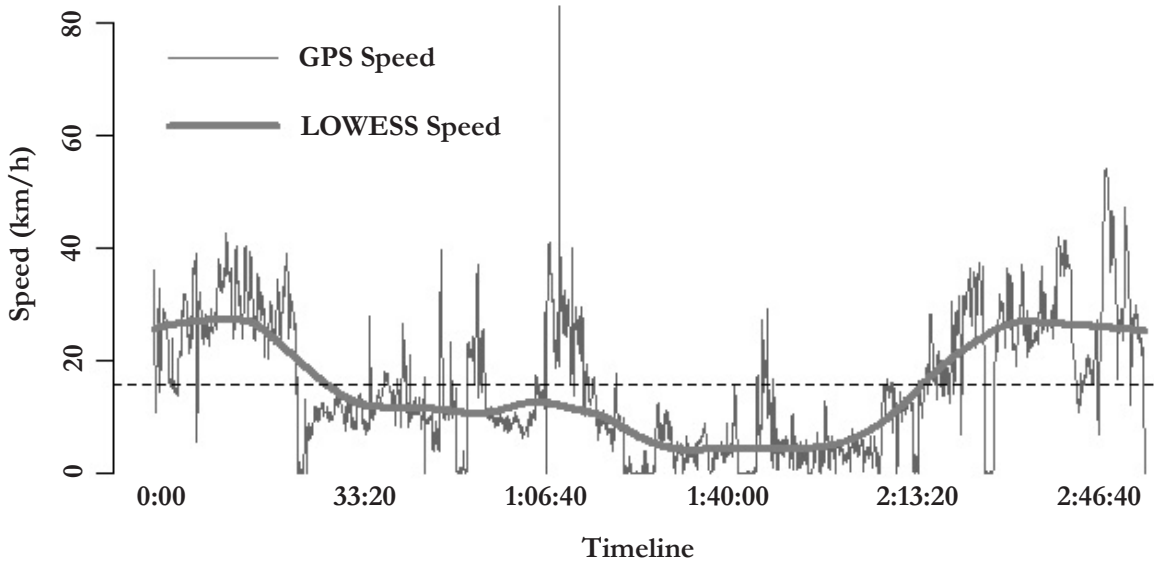
To have an insight into the interactive data map created with shinyapps.io please visit the following link:
https://dataviz19-noah.shinyapps.io/Garmin_Connect/

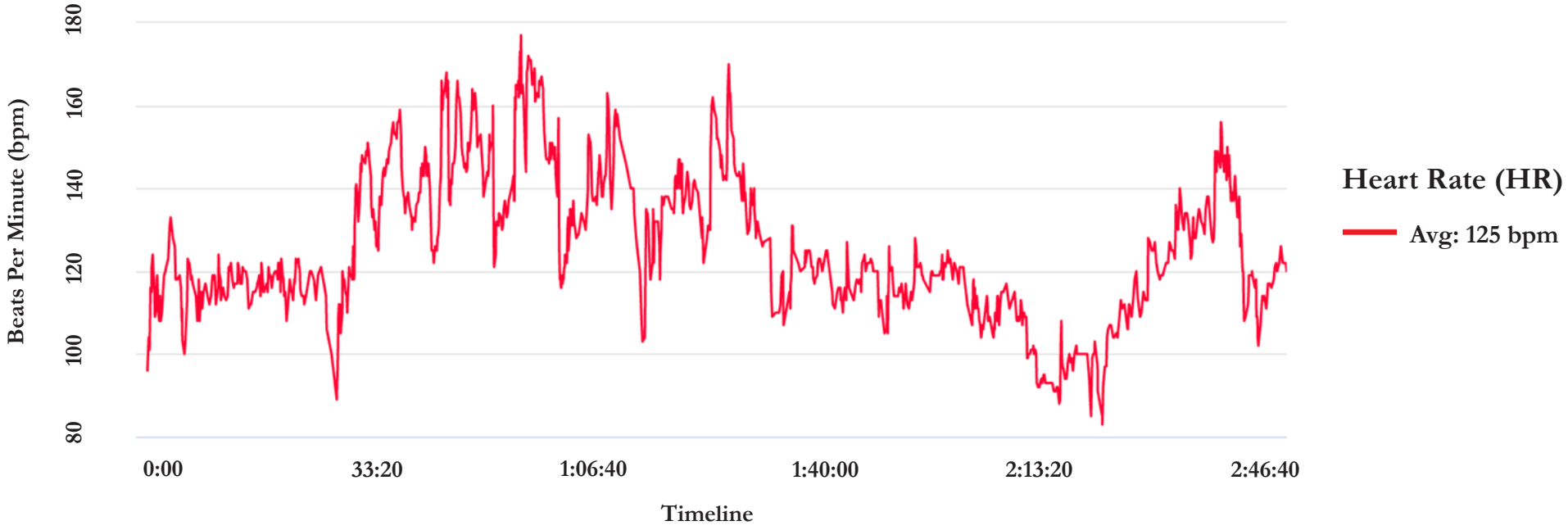


Garmin Activity Tracking

These graphs illustrate the physical data of my cycling route in Einsiedeln. The data points have again been generated with a Garmin Vivoactive HR and the dataset was provided from the official Garmin website.

While the upper plot is displaying my personal GPS and the LOWESS speed throughout the tour, the lower plot showcases my heart rate in beats per minute (bpm).

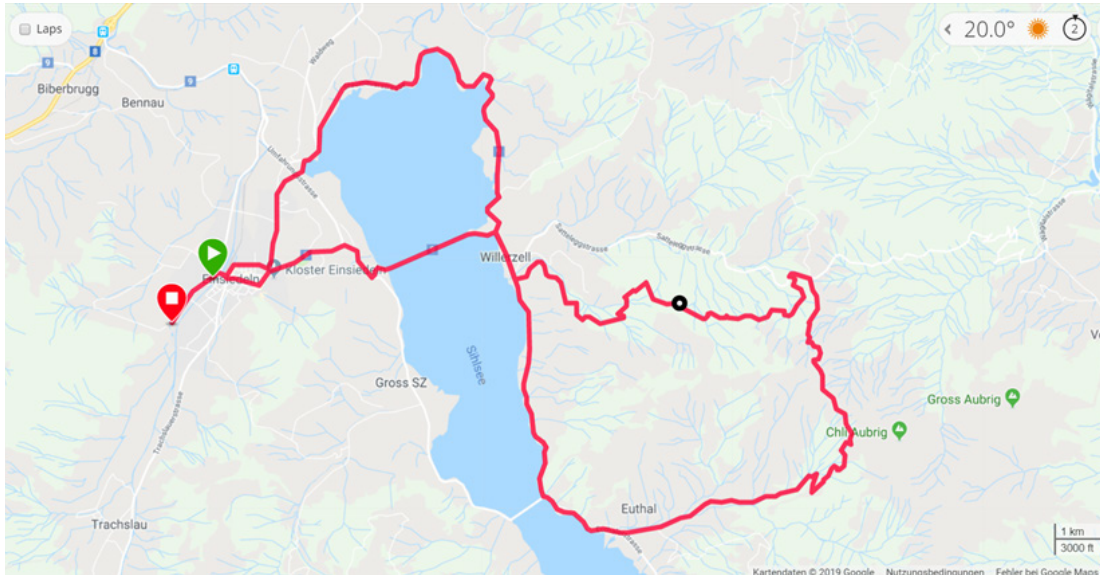


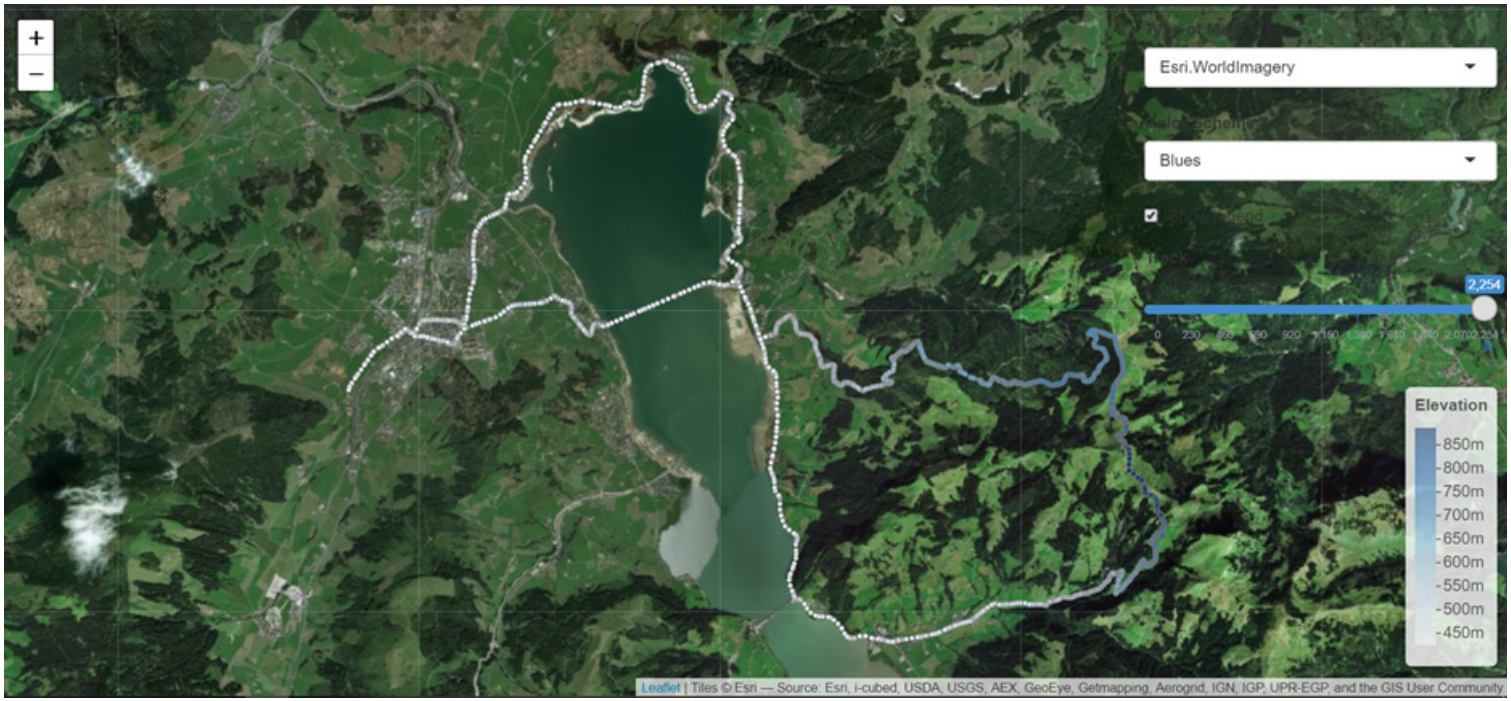


Garmin Activity Tracking

These maps represent the before shown graphs, providing further detail towards the tracking of this particular activity.

To have an insight into the interactive data map created with shinyapps.io please visit the following link:
https://dataviz19-noah.shinyapps.io/Garmin_Connect_Cycling/

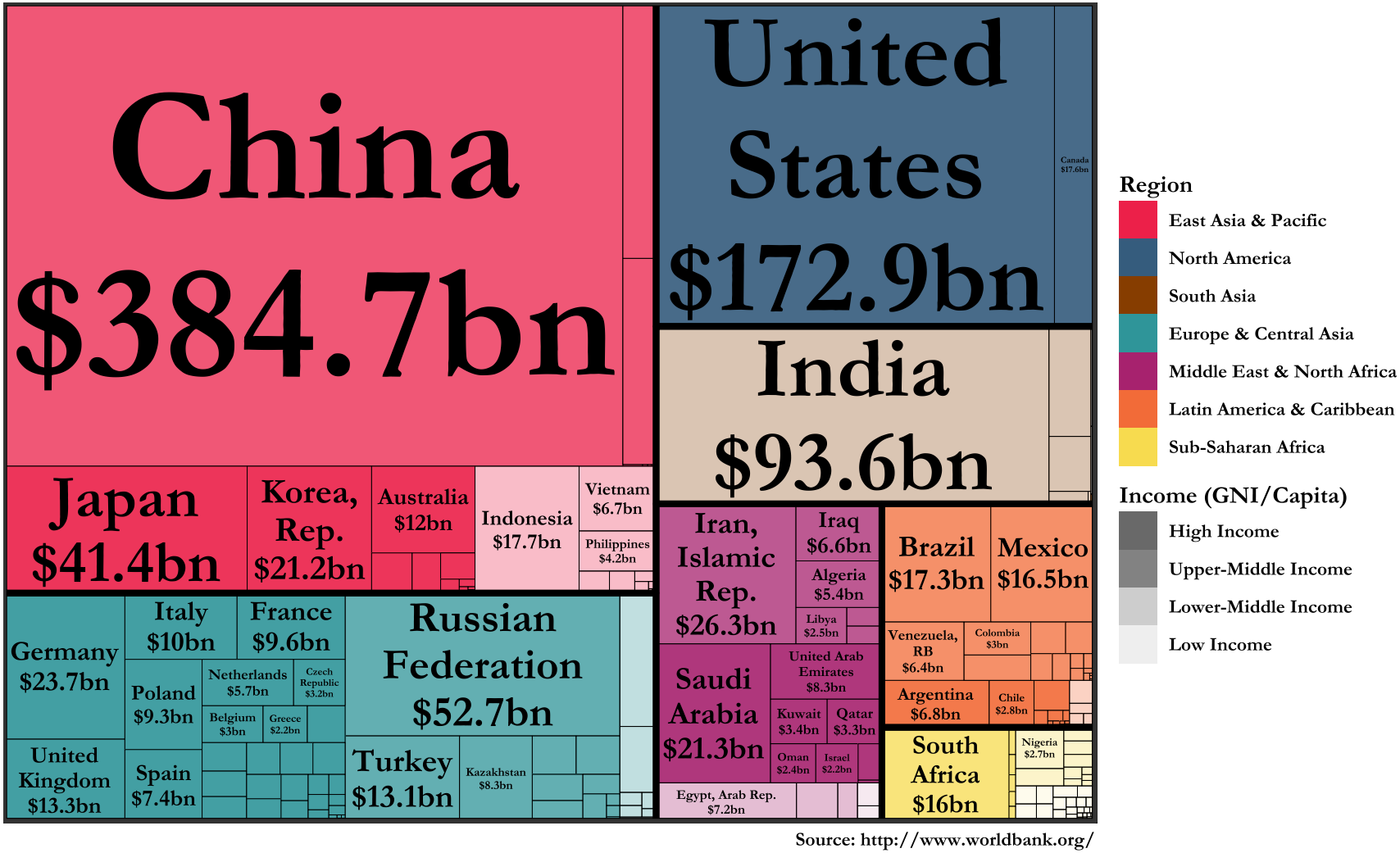




CO2 Emissions

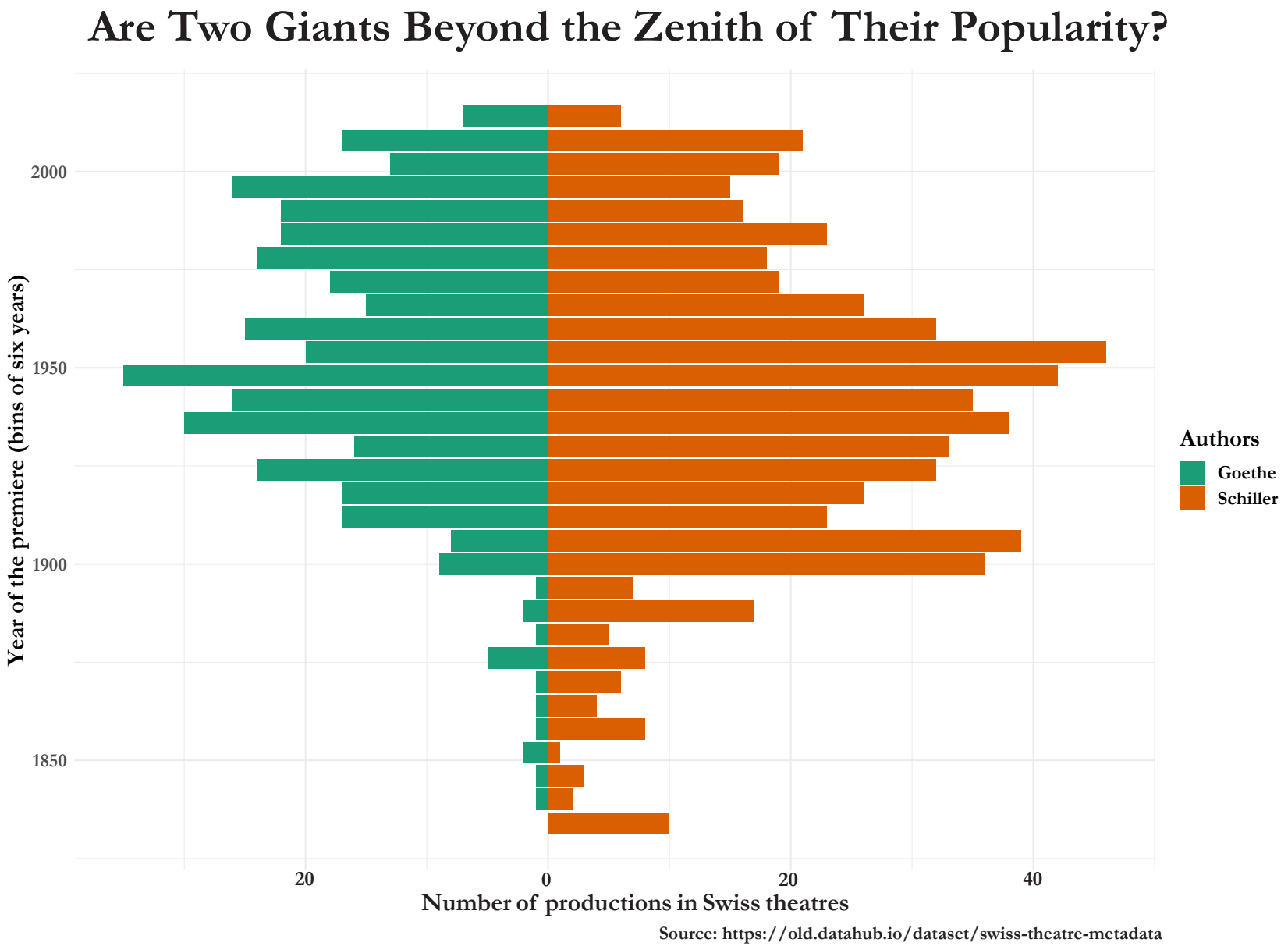
CO2 Damage in 2017 (US$30 per ton of CO2)

Most damage is caused by upper-middle and high income countries!

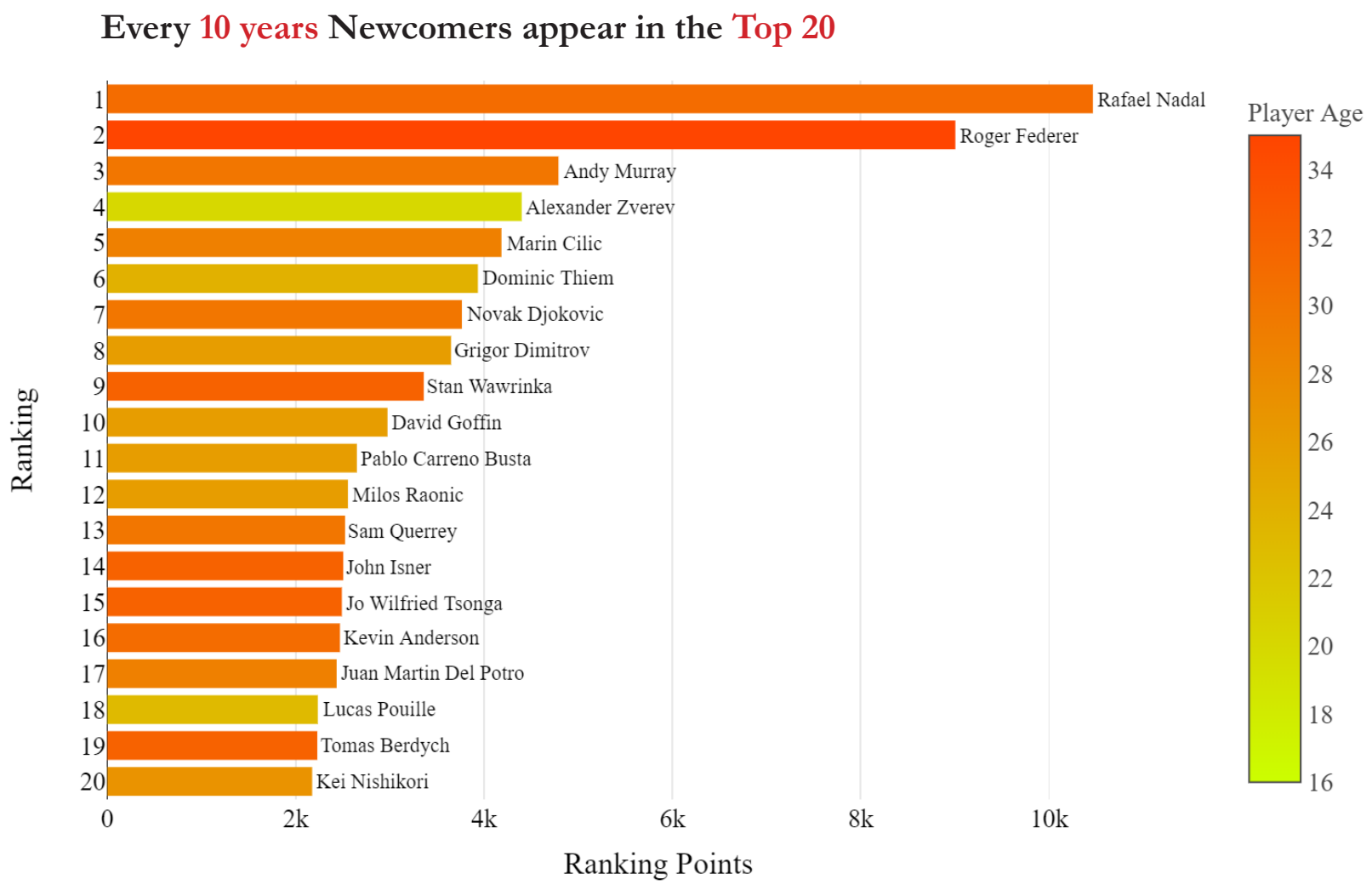


Source: http://www.worldbank.org/

Goethe vs. Schiller

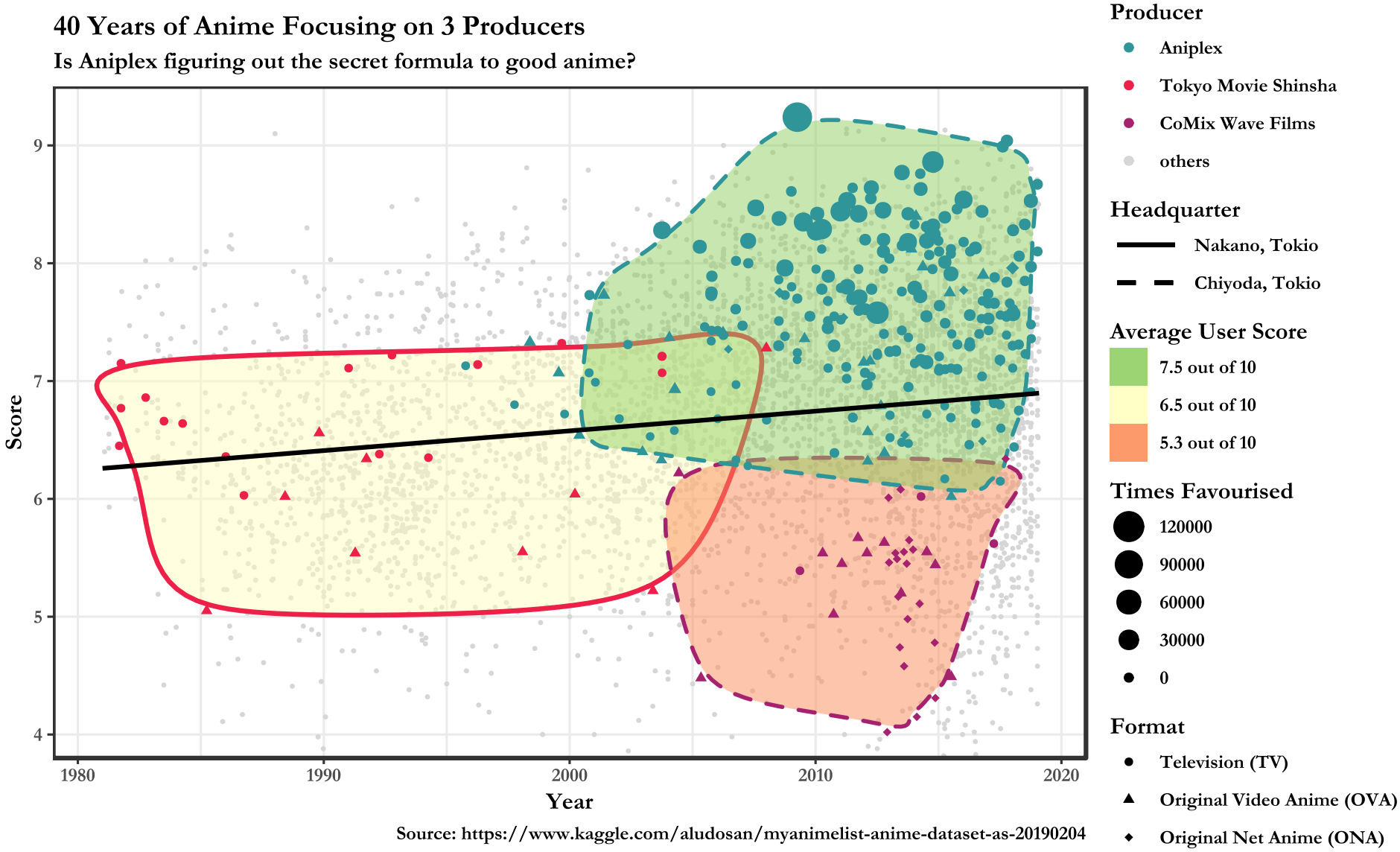


ATP Tennis Ranking



A shiny app always comes in handy for interactive plots and the respective one can be found here:
https://dataviz19-noah.shinyapps.io/ATP_Tennis_Ranking/

Anime Production



Favorite Tools

Shiny is a package in R Studio that allows users to create interactive applications that require a large amount of computational power. Further, these applications can be published on the Internet.

Adobe InDesign is a desktop publishing software application that can be used to create poster, flyers, brochures, etc. This portfolio was entirely assmbled and edited with InDesign and its toolbox.

Plotly maintains the fastest growing open-source visualization libraries for R, Python, and JavaScript. This feature can be easily implemented into your current project and allows creating beautiful plots.

Kaggle is the place to go when looking for datasets. It is owned by Google and allows users to find, publish data sets and much more. Many of the datasets that have been used for this portfolio where found there.

ggplot2 is a package in R Studio that assists its user in creating graphics, based on the “Grammar of Graphics”. After providing the data and how to map variables to aesthetics, it takes care of the details.

Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code. Jupyter was used for my personal viz, as it beautifully depicts the creation process.

Sources

Cover Page	Evolving Newsroom. (2016). Retrieved March 19, 2019 from https://evolvingnewsroom.co.nz/interactive-map-global-shipping-routes/
Student Data	Student.RData, provided by University of St. Gallen
Random Forest Performance on Testing Data	UCI. (2018). Retrieved April 20, 2019 from https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work
Prediction of Graduate Admissions	Kaggle. (2018). Retrieved March 25, 2019 from https://www.kaggle.com/mohansacharya/graduate-admissions Quote: https://www.goodreads.com/work/quotes/26567278-10-alone
Twitter Statistics	Sysomos. (2009). Retrieved March 28, 2019 from https://chandoo.org/wp/nightmarish-pie-charts/ Dataset: https://sysomos.com/inside-twitter/most-active-twitter-user-data/
Carbon Budget - Storytelling	Kaggle. (2018). Retrieved March 25, 2019 from https://www.nationalgeographic.com/environment/2018/12/climate-geoengineering-series-intro/
Spotify Data	Spotify. (2019). Retrieved April 30, 2019 from Spotify.com
Nobel Prize Winners	Kaggle. (2017). Retrieved May 03, 2019 from https://www.kaggle.com/nobelfoundation/nobel-laureates
Garmin Data	Garmin. (2019). Retrieved May 05, 2019 from Garmin.com
CO2 Emissions	World Bank. (2017). Retrieved March 22, 2019 from worldbank.org
Goethe vs. Schiller	Datahub. (2017). Retrieved March 21, 2019 from https://old.datahub.io/dataset/swiss-theatre-metadata
ATP Tennis Ranking	Datahub. (2018). Retrieved March 18, 2019 from https://datahub.io/sports-data/atp-world-tour-tennis-data
Anime Production	Kaggle. (2019). Retrieved April 23, 2019 from https://www.kaggle.com/aludosan/myanimelist-anime-dataset-as-20190204
Favorite Tools	Icons and Logos downloaded from various different websites
Extra Software	Jupyter Notebook, InDesign, Plotly

Grading

	Page(s)
Student data	1
Color as important aesthetics	2
Black-and-white viz	3
Bad viz	4
Improved viz	5
Good viz	6, 7
Viz about yourself	8
World bank data	15
Swiss data	16
Creation process	9, 10
Interactive viz	11
Data map	12
Viz with many aesthetics	18
Additional viz 1	13, 14
Additional viz 2	17
Favorite Tools	19

Collaborations	Page(s)
Jonas Roeser	2, 3, 11, 15, 16, 17, 18
Lukas Neuhauser	2, 16, 18
Samuel Halter	3, 15, 17