Proposal

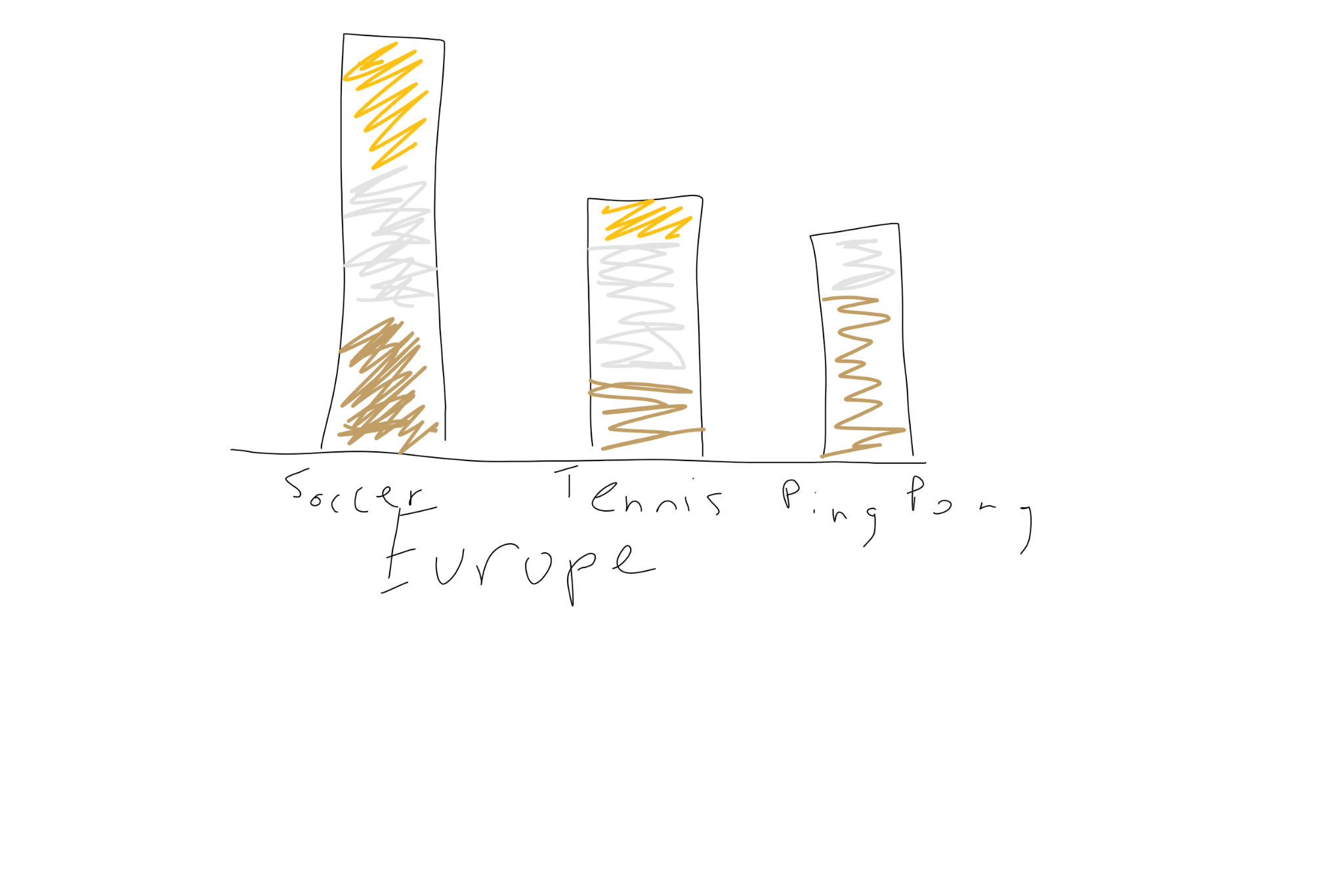
* Section 1 - Introduction: The introduction should introduce your general research question and your data (where it came from, how it was collected, what are the cases, what are the variables, etc.).
* Section 2 - Data: Place your data in the /data folder, and add dimensions and codebook to the README in that folder. Then print out the output of glimpse() or skim() of your data frame.
* Section 3 - Data analysis plan:
  + The outcome (response, Y) and predictor (explanatory, X) variables you will use to answer your question.
  + The comparison groups you will use, if applicable.
  + Very preliminary exploratory data analysis, including some summary statistics and visualizations, along with some explanation on how they help you learn more about your data. (You can add to these later as you work on your project.)
  + The statistical method(s) that you believe will be useful in answering your question(s). (You can update these later as you work on your project.)
  + What results from these specific statistical methods are needed to support your hypothesized answer?

**Project Proposal**

**Section 1:**

**Research Question: Which regions are the best in which specific sport?**

After initially discussing what route we wanted to go down with our project, scanning through many different websites such as Kaggle, the UK Gov data and the Harvard dataverse, we eventually stumbled upon a very cool dataset about Summer Olympic medal winners between the years of 1976 to 2008. The dataset has 11 variables, ranging from type of sport, type of medal won, country of athlete and many more. Our plan of action will involve breaking down the data into countries and eventually breaking it down into continents and which continents excel at which sports specifically.



**How are we going to analyze the data?**

An idea that we have is to possibly analyze each region then facet it by sport and fill by the medal colors. If we come into trouble with being able to see the graph, we can do the top ten sports, so it is easier to analyze the data. We have discussed and if it comes down to the top ten, we will pick as a group. We came to this conclusion because each region will have different sports that they like to stream during the Olympics. Our main goal is to see what regions excel at different sports. The region variable of our data, we are going to do by continents, but we know that Antarctica will not have any data, so technically there will be only six graphs.

We are going into it with assumptions such as North America will be the best at sports like Basketball, Africa will tend to be better at running, and Asia will be better at diving and ping pong. In general we know there are stand out countries such as Kenya in distance running or the USA in basketball, but it will also be interesting to look at regions as a whole, not just individual countries. We will be able to see this after grouping by continent and creating a graph that differentiates between which medals were won in which sport. We could also buttress these graphs with something like correlation statistics, we will see what works best for the data when we start manipulating it.

**Section 2:**

Data: <https://www.kaggle.com/divyansh22/summer-olympics-medals>

Section 3:

The variables we will be using are medals, region (which we will create), country and possibly a few more. We are looking to see which regions of the world are the best at which sports, and how they differ around the world. We are looking to use summary statistics as well as geom\_bar and geom\_point to show how certain regions excel at certain sports. We are excited to get to work on this project and see how the project plays out and what findings we have.