CSE 163 Report

Jenny Chen, Luotong Kang

Summary of Questions and Results

1. In which region do people have the highest heights for both male and female, and which region appears to have the lowest in 2022? We want to graph out the height data on the map and find out the region with color that describes the highest value of height data. Vice versa, the same map is used to observe for color that represents the lowest height data.

   Result - From the color key, people from European countries especially the region of northern Europe have the highest heights for both male and females.

2. What is the relation between height, GDP per capita, and time? Use mean height information for 18-year-old boys from the year 2009 to 2019 in different countries. Explore the correlation between GDP per capita of countries and mean height. Broad understanding and elimination of distractions are the main goals of this step.

   Result - There exists a correlation between GDP per capita of countries and mean height of countries. But the height change is so small over time for countries with low GDP per capita, we cannot clearly conclude the correlation.

3. How does GDP per capita impact the overall heights of adolescents with age 18 in specific countries in the year 2019?

Result - The correlation between the two factors is very weak from the graph, so we cannot conclude that GDP per capita has an impact on the heights of 18-year-old adolescents.

4. What are the trends of height growth in China and in the U.S. between 2005 to 2019 (matching with age 5 to age 19)? How does that tell us the general tendency of growth (height difference between 5 to 19 year old people)?

   Result - The shape of the two graphs is mostly the same, implying that the growth trend is the same for both countries. We suggest that country difference is not a factor for growth trends. Still, we cannot rule out the possibility that the variation happens before age 5.

Motivation

Based on the heights given in the dataset, we are interested to know about if heights can be varied due to geographical regions in both genders. Also, we want to know if the trends of growing in heights (e.g., comparing the growth trend from age 5 to 19) can be different based on country bases.

We noticed from the dataset that people from specific regions tend to have the highest average heights while people from some regions tend to have the lowest average. Thus, we wonder if heights and income levels are correlated factors since we've observed some regional patterns. Income becomes one of the factors that we want to consider since income level can decide nutrition intake, which is an important element for one's height.

Our group members both come from China and attend college in the U.S., and this makes us interested in comparing the trend of height growth in the two countries.

Dataset

1. This dataset includes data categories of country, sex (boys and girls), year (1985 - 2019), age group (5 - 19), mean height, mean height lower 95% uncertainty interval, mean height higher 95% uncertainty interval, and mean height standard error.

   https://www.ncdrisc.org/downloads/bmi_height/height/all_countries/NCD_RisC_Lancet_2020_height_child_adolescent_country.csv

2. This dataset provides data of male/female height by countries in the year of 2022.

   https://www.kaggle.com/majyhain/height-of-male-and-female-by-country-2022?select=Height+of+Male+and+Female+by+Country+2022.csv

3. This dataset from the World Bank provides GDP per capita by countries for all years with records.

   https://api.worldbank.org/v2/en/indicator/NY.GDP.PCAP.CD?downloadformat=csv

4. Used built-in dataset "naturalearth_lowres" as cartographic information.

Method

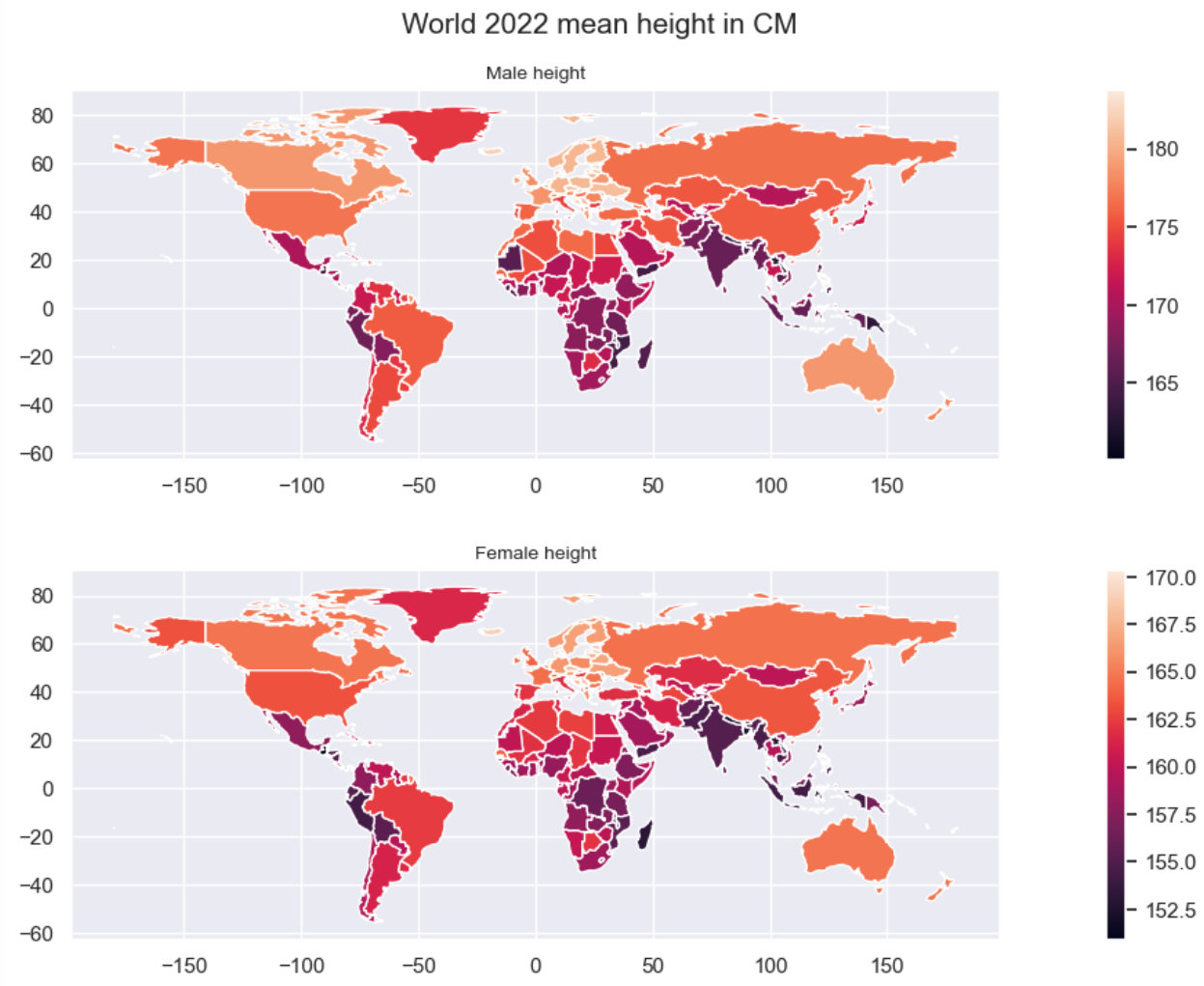
1. Load necessary libraries and read data. Use built-in GeoDataFrame "naturalearth_lowres" as the base of map visualization. There's an inconsistency of country names, which causes us to start with a slow pace. Strategic contours are found from library `dataprep`, in which we convert all country names into alpha-3 format as a reference column. Drop all entries with undefined alpha-3 names, and manually drop areas with duplicated alpha-3 names (Macao SAR and Hong Kong SAR). Done in a separated script.

2. For research question 1, merge (inner-join) the existing map GeoDataFrame with 2022 mean height DataFrame using alpha-3 country names. Plot that GeoDataFrame data directly using matplotlib (one for male and one for female), with legend as colorbar.
3. For research question 2, we use a separate python file to store function because of complexity. First we need to melt the GDP dataframe to make multiple columns of GDP per capita by years to be melted into one column and an additional year indicator column. Then filter and merge with age 18 height information to get a graphing DataFrame. Create a model using SVR from sklearn and do regression. Create a meshgrid and plot using plotly. Save interactive graph as html. Use a screenshot in the report.
4. For research question 3, preprocess the two DataFrame to only include 2019 data. The first graph is generated by adding the 2019 GDP per capita column to GeoDataFrame and plotting as a map. Filter to only include countries with GDP per capita < 13000. The second graph is generated by adding the 2019 GDP per capita column to filtered (only keep age 18 in 2019) mean height dataframe and plotting using seaborn lmplot. The third graph is generated by adding the 2019 GDP per capita column to filtered (only keep age 18, but in multiple years) mean height dataframe and plotting using seaborn scatterplot.
5. For research question 4 part 1, pick out height data for China and the United States for the generation born in 2000. This is to say, keep Age 5 for data from Year 2005, Age 6 for data from Year 2006, etc.. Do the same to include height of people born in 1980. Graph using relplot to compare countries, generations, and sex differences.
6. For research question 4 part 2, first pivot age group to create columns of age 19 and age 5 for each born year and each country. Then calculate growth (mean height age 19 - mean

height age 5 for boys. Visualize the mean growth by GDP per capita by setting this discrete regression using regplot and x_estimator.

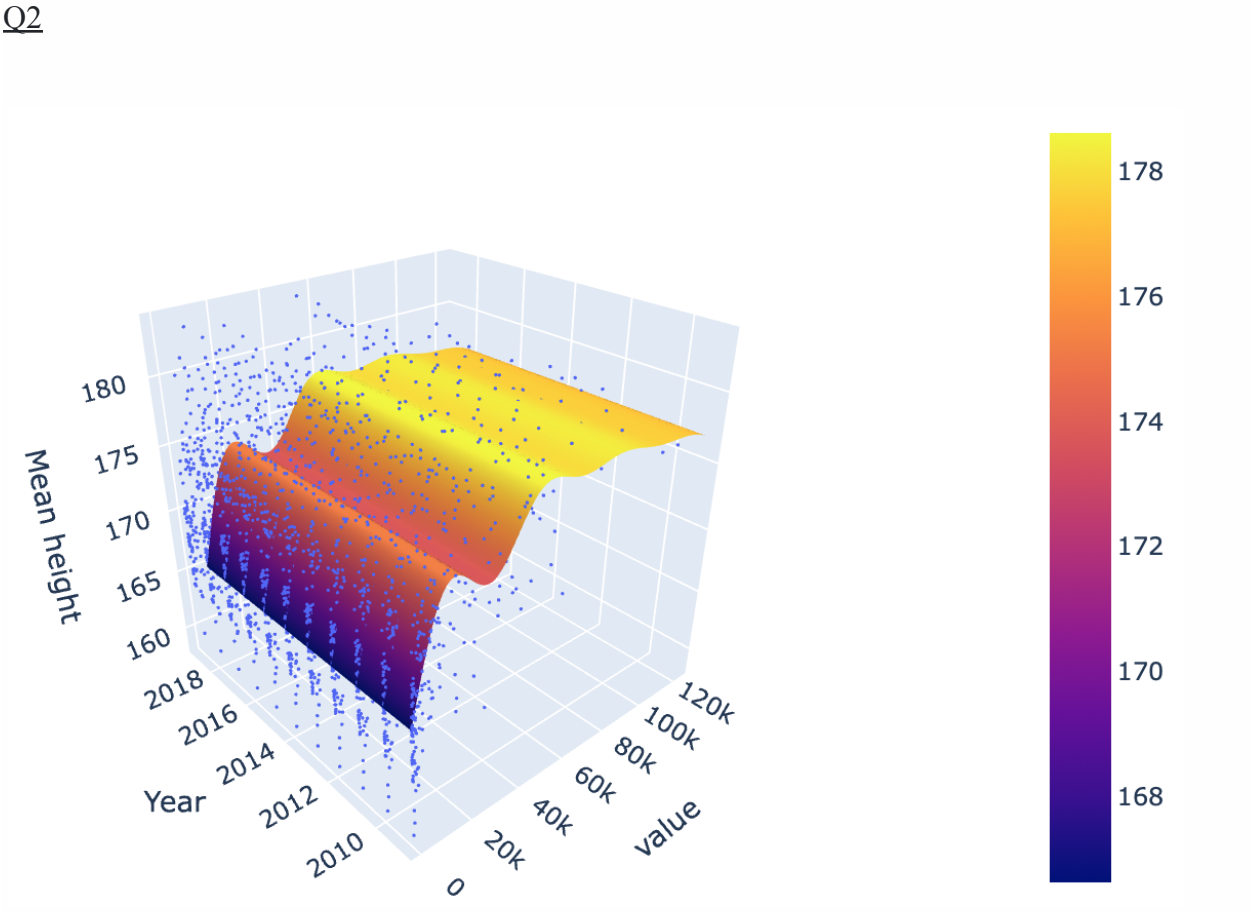
Results

Q1



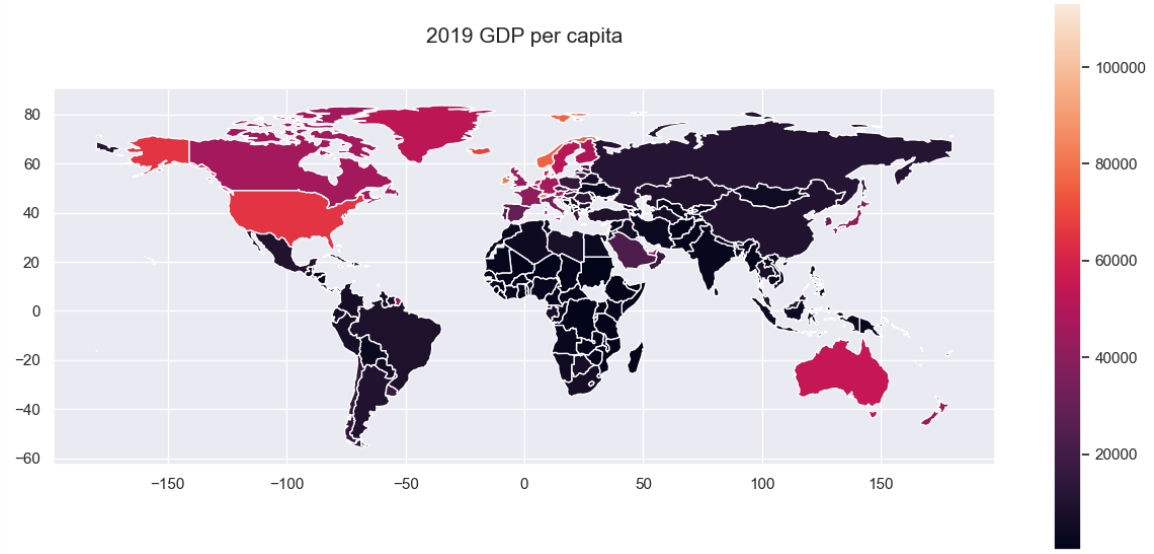
From this graph, we can notice that countries near the equator typically have lower mean height, while countries in high latitude tend to have higher mean height. Europeans are outstanding in heights, while Africans sometimes have low mean height that cannot be predicted using latitude. Relative ranking is the same for males and females.

Q2



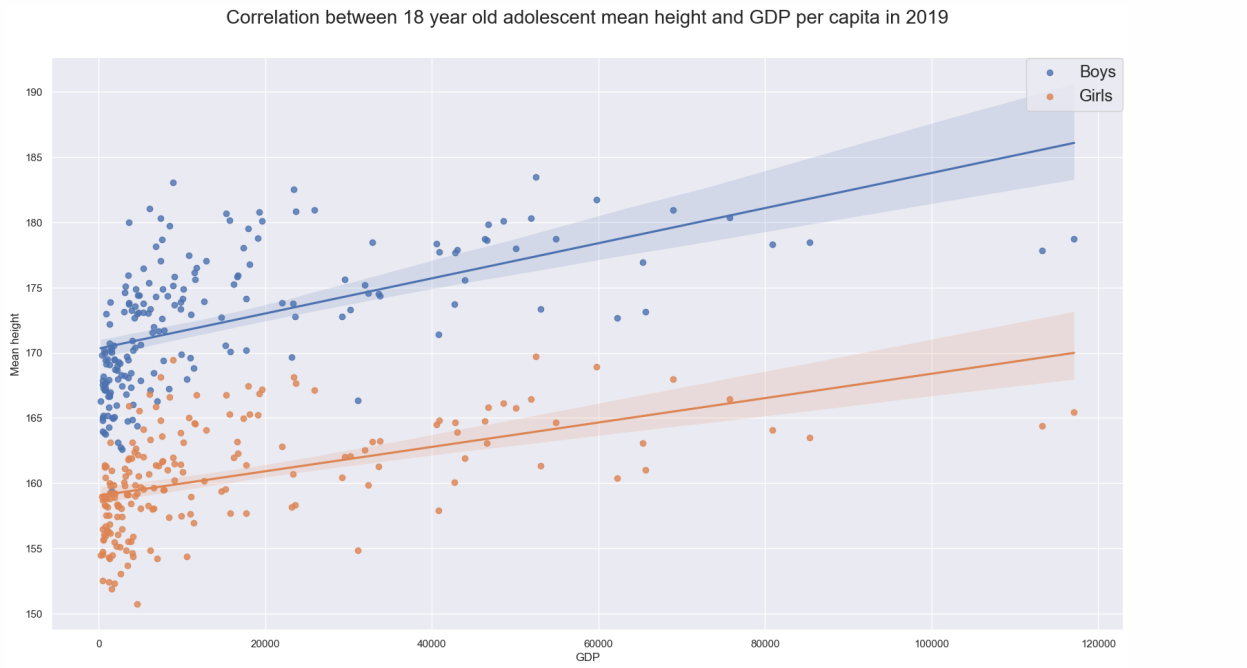
In the graph, “value” axis represents the measure of GDP per capita. The inserted picture in static format is not an effective visualization of data. Please refer the following to the html file in the same folder. On one hand, the graph shows that there might exist a positive correlation between GDP per capita and mean height of countries. On the other hand, it shows trivial height change over time for countries in low GDP per capita.

Q3

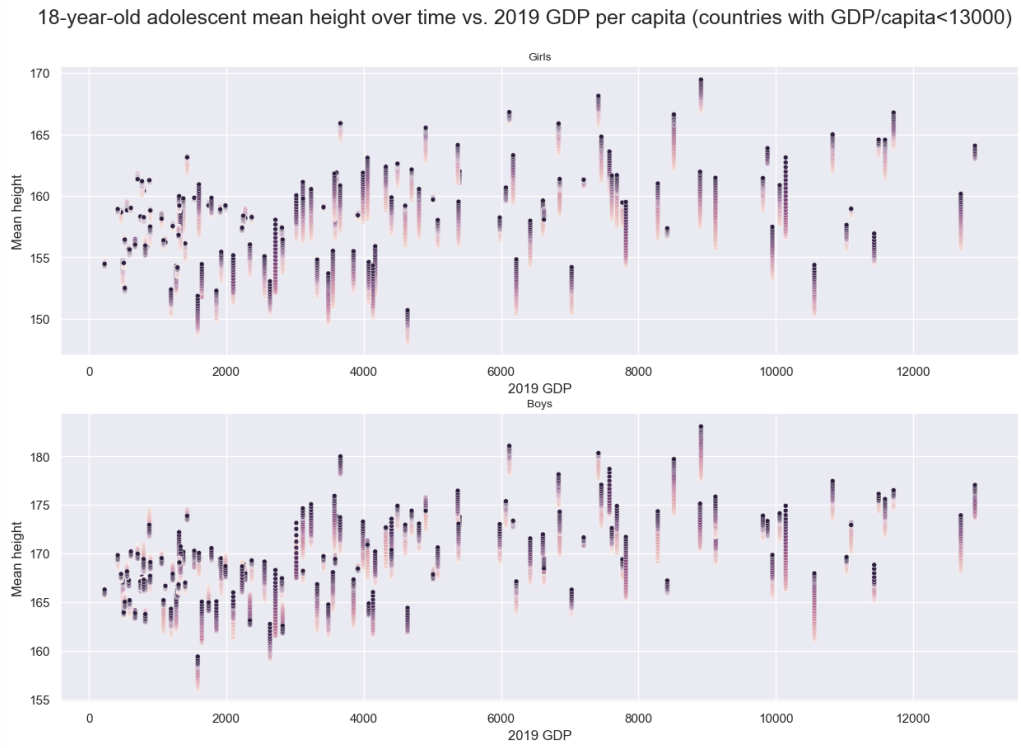


After graphing out this map, we decided to focus on countries with GDP per capita under 13000 for the rest of question 3 for the following reasons:

- Countries with low population and an enough economic foundation will not reflect the influence of height by income level. E.g. Bermuda now has the fourth highest per capita incomes in the world, primarily fueled by offshore financial services for non-resident firms, especially offshore insurance and reinsurance, and tourism. Population 63,903
- Small reference group: the regression won't perform well in sparse graph.



This regression graph validates our guess that GDP per capita and mean height is positively correlated for countries with GDP per capita < 13000. Also, there's a clear difference between sex that matches our intuition. Another visualization below that give some hint about question 4:



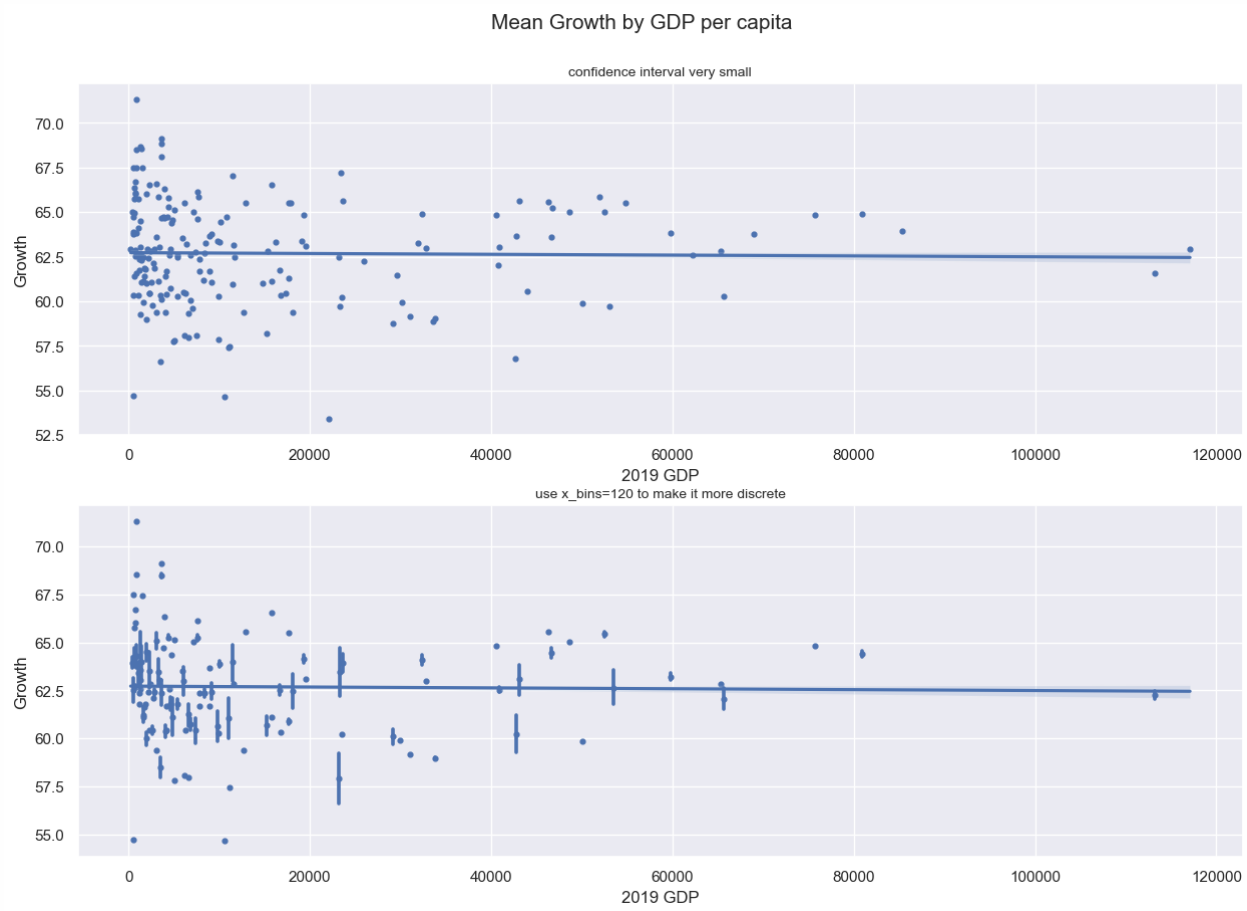
Q4

From the graph, we can see that there's no obvious change for US adolescent height from the generation born in 1980 to those who were born in 2000. However, there's a clear increase in Chinese adolescent height from 1980 born people and 2000 born people. In addition, from the parallel relationship of lines, we come up with some guesses:

1. Change of height between generations is not related to the mean height change from 5 to 18 years old (let's call it "growth") of a group of people. This means that the variation between generations is happening before age 5.
2. This growth seems unrelated to country differences.

We'll validate our guess by doing a linear regression to growth. Since there's clearly a difference in growth by gender, we'll choose to only look at boy's data. Each point in the graph will show the mean growth across generations (average the difference) of one country. All countries are lined up by GDP per capita horizontally. The regression will show GDP vs. mean

growth across generations. Hypothesis 1 is supported if the confidence interval of most points is not wide. Hypothesis 2 is supported if the regression line seems horizontal.



The upper graph actually is graphed with confidence intervals of the mean for each points, but the interval is too small to be reflected on the graph (which strongly supports our hypothesis. In order to show the validity, we plot the second graph by making it more discrete (use x_bin). We can see some small intervals here. It appears that the regression is nearly horizontal, which proves the soundness of our hypothesis that height growth is not related to GDP and the ultimate difference in height is likely to be determined before age 5.

Impact and Limitations

Our analysis shows that GDP per capita is positively correlated with height. However, there's so many limitations to draw a conclusion. The cause & effect relationship is doubtful. First of all, GDP per capita is insufficient in indicating life quality. For example, high total GDP countries with high population tend to have a better life quality compared with what the model would tell us. The variation of races and ethnicities can also be huge factors that affect the height growth no matter how the nurturing process after birth changed by income level. Many other compounding factors affect people's height. In addition, our analysis doesn't consider the initial standard deviation given by the dataset.

Wrong implications of our results include concluding biasly that people from certain countries must be very tall/short, prefering one sex over another given the height difference when selecting candidates for certain jobs, etc. Potential bias may also exist in our data because of the missing data. Inferencing conditions to small countries and controversial areas ought to be very careful.

However, research question 4 does give us some interesting results. The tiny change of growth (age 19 mean height - age 5 mean height) over time seems to be a general trends for all countries. It tells us that our ultimate height is determined by height at age 5! In terms of what this report can teach us, I would say that nutrition before age 5 plays a very important role in growing tall.

Challenge Goals

Multiple Datasets

We used multiple datasets adding on to the heights dataset in order to analyze correlations between heights and other factors. Data such as GDP serves as supplementary data to complete the analysis in our research questions.

New libraries

We applied a new library not learned in class in our project to further improve the visualization effects of our graphs and images. We used Plotly to visualize Question 2.

Work Plan Evaluation

Plan	Expected Time	Actual time used
Data trimming and cleaning	1 hour	3 hours
Data manipulation	2 hours	2 hours
Data Visualization	4 hours	4 hours
Testing	1 hour	1 hour
Writing report	4 hours	4 hours
Making video	2 hours	N/A

1. Data trimming and cleaning

   We want to sort out data and leave rows or columns which we need for our research questions.

   Evaluation: We thought that cleaning process is simple, but it appears that the inconsistent in country names is a big trouble. We spent a lot of time and finally solved it by using a new library.

2. Data manipulation
   a. Q1: Merging the height dataset from 2022 with map dataset.
   b. Q2: Melting height dataset in different age and merging with GDP information.
   c. Q3: Filter the GDP per capita to only include 2019. Joining the height dataset from 1985 to 2019 and the filtered dataset.
   d. Q4: Calculate a growth column using the dataset from 1985 to 2019. Extract data using born year.

   Evaluation: The time used here is within our expectation. Although we didn't plan for Question 4 in proposal, the desirable cleaning technique found in previous step lighten the burden.

3. Data visualization
   a. Q1: directly plotting the merged GeoDataFrame and look for the highest and lowest by looking at the color of each region.
   b. Q2: use plotly.
   c. Q3: use both map plotting technique and plots from seaborn .

d. Q4: use relplot and regplot from seaborn.

Evaluation: The time used here is within our expectation. 3D visualization is as time-consuming as we expected. Adjustments of styling (legend, font, etc.) are also foreseen when we wrote the proposal.

4. Testing
    a. Test the cleaning technique from package dataprep.
    b. Test the existence of confidence intervals in second visualization of question 4.

Evaluation: the time used here is less our expectation. Since we do not know what need to be tested when writing proposal, the estimate time is inaccurate. Given that we are not doing Machine Learning, the testing process is relatively simple.

5. Writing a report

Finalize the result and put the results in a presentable format.

Evaluation: The time used here is within our expectation. Co-working functionality of Google doc is helpful.

6. Making videos

Make slides that contain visualizations and result outlines. Use zoom to record.

Evaluation: We haven't made our videos when writing this report. But we expect the time to match with the estimation in proposal.

Testing

1. Shows the function “clean_country” from package dataprep works properly. The function should convert various expressions of countries into standard alpha-3 names. For example, the function converts both “US” and “United States of America” into “USA”.
2. In the second visualization of question 4, confidence intervals for each dot should exist. We’ll prove that our graph does plot out the interval, so that our argument of them being very small is valid. Test by comparing the max growth with the mean growth of selected country. We expect the comparison to return boolean False, and assert_equals() gives us no error.

Collaboration

Dataprep documentation: https://docs.dataprep.ai/user_guide/clean/clean_country.html

Plotly documentation: https://plotly.com/python/ml-regression/#basic-linear-regression-plots