In [ ]:
```
# Libraries
import numpy as np
import pandas as pd
import altair as alt
```

PSTAT 100 Project plan

Group information

Group members:

Yibo Liang, **Aarya Kulkarni**, **Alan Su**, and **Nicole Magallanes Flores**

Contributions:

1. Alan Su: worked on finding data and data background (Part 0)
2. Yibo Liang: worked on tidying the dataset/initial exploration (Part 2)
3. Aarya Kulkarni: worked on data description (Part 1)
4. Nicole Magallanes Flores: worked on planned work (Part 3)

Background

Education is an important issue in every country around the world, as it is with education that people can innovate and drive their country forward. But while education is obviously important to many people, not everyone is able to have access to an education. Some countries may not have the resources to put every citizen through school, or some countries may have the resources, but for example, people that live in rural areas might choose not to go through education in favor of continuing a family business or pursuing their own economic endeavors.

The dataset that we have chosen to work with is a dataset taken from The World Bank's data catalog, their dataset on educational attainment and enrollment around the world. This dataset takes data from many types of surveys from all around the world in order to get data on things such as socioeconomic conditions, or living standards such as rural or urban. Using this data, we will explore how educational attainment is affected by conditions such as wealth, gender, and living location, and can use this information to advocate for better educational attainment standards everywhere in the world. Is the world bank right when they say that there are large differences in educational attainment by wealth? Will gender and living standards affect a person's educational attainment? Using this data, we plan on looking into these questions and hope to gain a better understanding of what educational attainment is like around the world.

---

1. Data description

The data describes the percentage of population ages 15 to 19 that has completed each grade (1-9) in developing countries around the world spanning multiple years (1990-2020).

This data came from a World Bank database with summary information on education level attained, taken from household surveys from developing countries around the world (http://www.worldbank.org/en/research/brief/edattain). Household-based surveys used to create this dataset include: Demographic and Health Surveys (DHS- http://www.measuredhs.com), Multiple Indicator Cluster Surveys (MICS- http://www.childinfo.org), Living Standards Measurements Study Surveys (LSMS- http://www.worldbank.org/lsms), and other household-based surveys (country specific), ex: socio-economic surveys. In addition to these household-based surveys, selected country/year variables were added to the dataset from the World Development Indicators database (http://databank.worldbank.org). Some variables added include gdp per capita (based on 2015 U.S. dollars) and the consumer price index (based on 2010) for various countries.

The relevant population is the population ages 15-19 in the countries surveyed. Since the dataset is aggregated from multiple surveys. Some documentation was provided on two of the types of household surveys performed.

The Demographic and Health Surveys (DHS) used a mix of questionnaires, biomarkers, and geographic information as survey tools to conduct this survey. This survey's sampling design consisted of a two-stage stratified cluster design in which Enumeration Areas (areas canvased) were drawn from Census files (stage 1), followed by a sample of households being drawn from a list in each Enumeration Area (stage 2).

The Living Standards Measurement Study Surveys (LSMS)'s sample frame is given by the Population and Housing Census. Following this, a two round sampling method is used. The first round selects Primary Selection Units (PSU) through random sampling. The second round selects subunits from each PSU using a method of systematic sampling.

Since the sampling frame is usually given by the country's Census, and the sampling mechanism involved random sampling, we can say that the scope of inference are the households captured in the Census.

Data semantics and structure

Units and observations: The observational units are 82 unique countries around the world from 1990 to 2020.

Variable descriptions:

Name	Variable description	Type	Units of measurement
country	observation country	Character	Name of Country
year	year of observation	Numeric	Calendar year
gdppc2015	gdp per capita (2015 U.S. dollars)	Numeric	gdppc
cpi2010	consumer price index (Based on 2010)	Numeric	cpi(hundreds)

Name	Variable description	Type	Units of measurement
level	level of education	Numeric	grade
sex	sex of the group of children aged 15-19	Factor	male/female
location	the location of the group of children aged 15-19	Factor	urban/rural
prop	proportion of the group of children aged 15-19 that attained an education	Numeric	percentage

Example rows:

In [ ]:
```
# load tidied data and print rows
data_tidy = pd.read_csv('../data/tidy_data.csv').drop(columns='Unnamed: 0')
data_tidy.head(5)
```

Out[ ]:

	country	year	gdppc2015	cpi2010	location	sex	level	prop
0	Afghanistan	2015	556.007	132.883	Urban	Male	1	0.926
1	Afghanistan	2007	392.710	83.074	Urban	Male	1	0.846
2	Afghanistan	2010	526.104	100.000	Urban	Male	1	0.862
3	Angola	2015	4166.980	159.405	Urban	Male	1	0.976
4	Angola	1999	2458.096	0.684	Urban	Male	1	NaN

---

2. Initial explorations

Basic properties of the dataset

Dimensions:

In [ ]:
```
data_tidy.shape
```

Out[ ]:
```
(22536, 8)
```

The tidy data has 22536 rows and 8 columns.

Missing Data:

In [ ]:
```
data_tidy.isna().mean()
```

Out[ ]:
```
country      0.000000
year         0.000000
gdppc2015    0.011182
cpi2010      0.070288
location     0.000000
sex          0.000000
level        0.000000
prop         0.015974
dtype: float64
```

There are so missingness in our tidy data. Specifically, `gdppc2015` is missing about 1.1% of values, `value` is missing about 1.6% of its values, and `cpi2010` is missing about 7% of its values. This data is missing because the country did not report its gdp/cpi that year.

One interesting thing of note is that year is not continuous for all countries. Some countries did not report anything for certain years; therefore, we need to keep track of these gap years.

Variable Summaries:

Alot of our variables are binary factors; so we will take a look at the most important variables of interest: `gdppc2015`, `cpi2010`, and `prop`.

Ideally, we would group these variables by years; however, we will not do so here for the sake of space, but keep in mind that the rest of the work will consider these variables in the respective year groups.

Below are the count, mean, std, min, 25%, 50%, 75%, and max statistics for `gdppc2015`, `cpi2010`, and `prop`:

In [ ]:
```
data_tidy.loc[:, ['gdppc2015', 'cpi2010', 'prop']].describe()
```

Out[ ]:

	gdppc2015	cpi2010	prop
count	22284.000000	20952.000000	22176.000000
mean	2570.711876	82.319928	0.713313
std	2597.316331	48.889600	0.282070
min	236.461000	0.000000	0.000000
25%	751.473000	51.094000	0.526000
50%	1579.700000	79.160500	0.816500
75%	3674.354000	106.762000	0.955000
max	28693.063000	508.339000	1.000000

The summary statistic alone is not enough to make initial generalizations about our data. As mentioned earlier, we should group the data by `year` and analyze the statistics of each `year`. We will briefly look at the statistics with consideration to year below in the `Exploratory Analysis` section.

Exploratory analysis

In [ ]:
```
alt.data_transformers.enable('default', max_rows=None)
from visualize import line_plot1, line_plot2, line_plot3, line_plot4
```

Visualize Trends

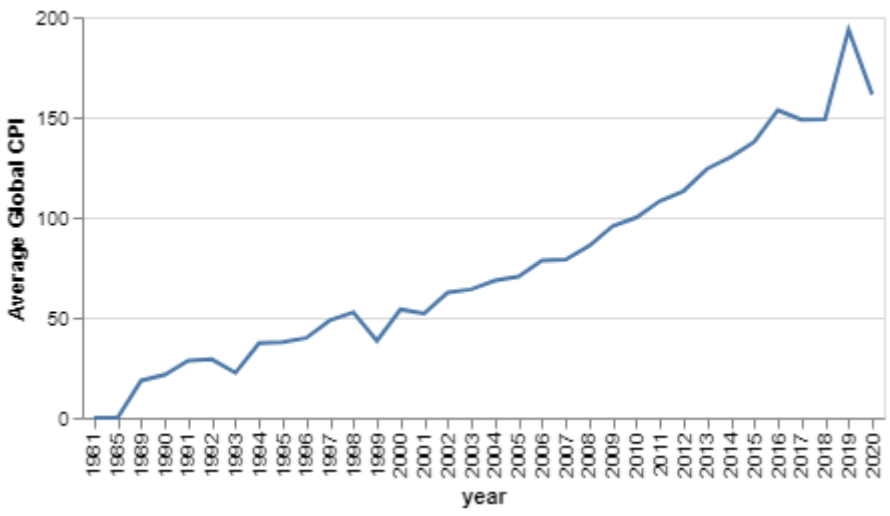
We won't do any complex EDA here, but we will take a look at some trends of our statistics when grouped by `year` through some simple line charts.

1) `Average CPI` against `year`

The following is the average `cpi2010` for all countries per year. Clearly, there is some positve correlation between `cpi2010` against `year`.

In [ ]:
```
line_plot1(data_tidy)
```

Out[ ]:

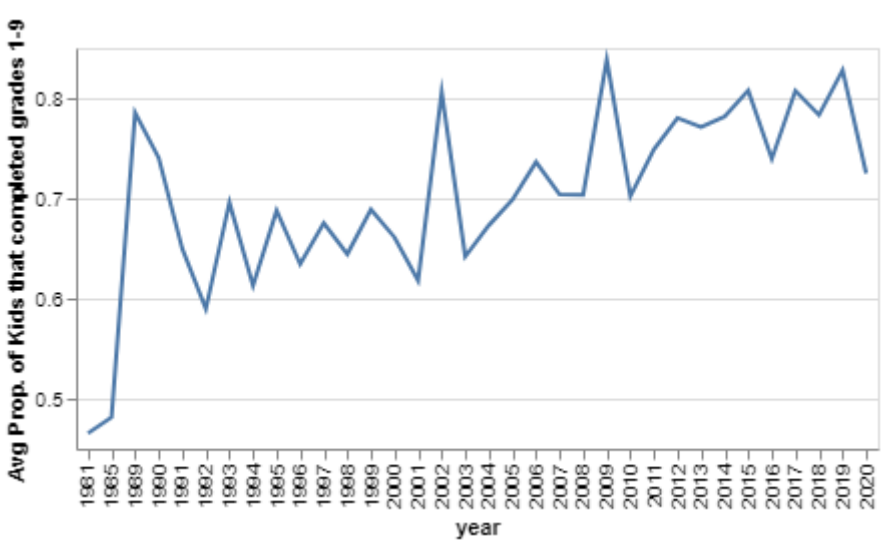


2) `Average prop of kids that completed grades(1-9)` against `year`

Again the chart below show some relationship between `prop` against `year`.

In [ ]:
```
line_plot2(data_tidy)
```

Out[ ]:



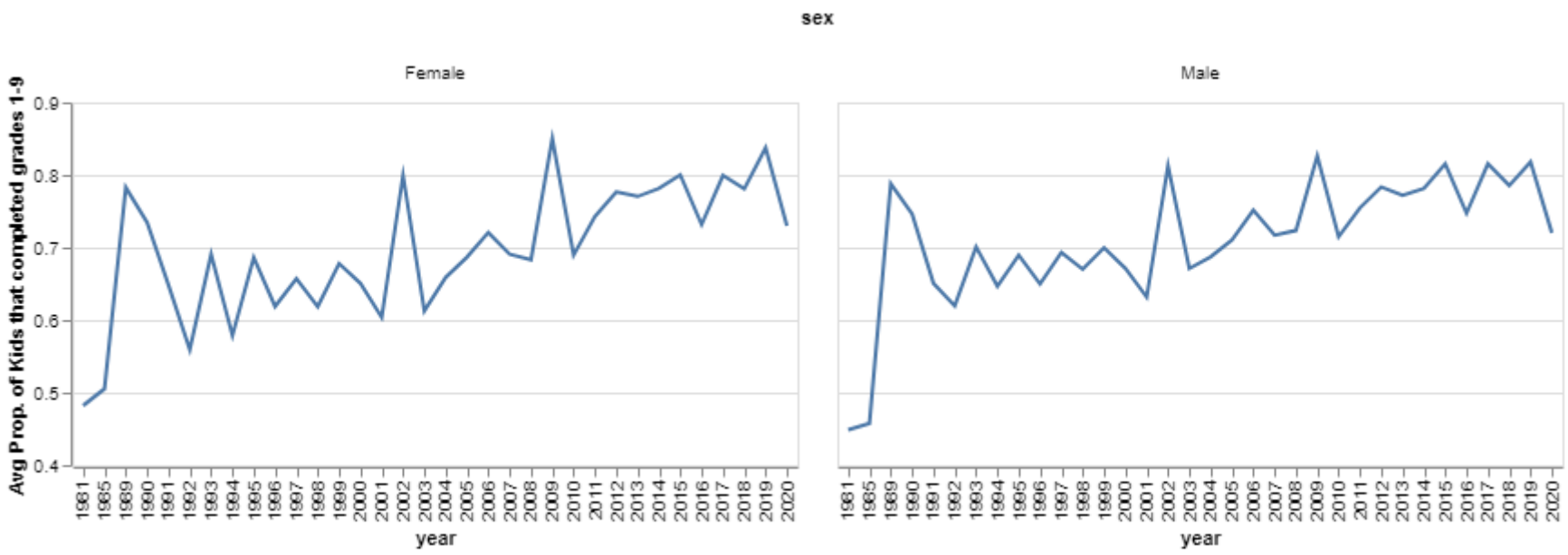
The natural followup question to plot-2 is whether `sex` and `location` of the kids affect the `prop`.

Below are two plots of plot-2 faceted by `sex` and `location` respectively:

3) `plot2` faceted by `sex`

In [ ]:
```
line_plot3(data_tidy)
```

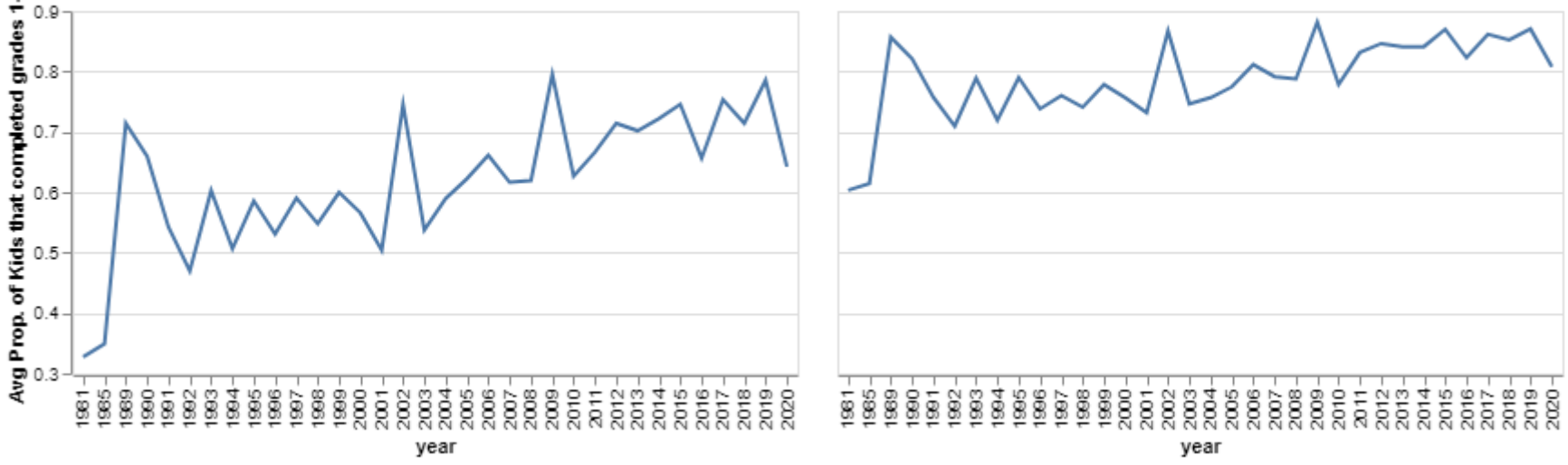
Out[ ]:



4) `plot2` faeted by `location`

In [ ]:
```
line_plot4(data_tidy)
```

Out[ ]:



It's hard to tell the difference between Male and Female with this simple line chart, but we will explore their effects further in part two. On the other hand, it looks like kids in Urban areas retain a higher `prop` overall when compared to kids in Rural areas; we will explore this further in part two also.

---

3. Planned work

Questions

1. Are there significant differences in educational attainment depending on wealth?
2. Do gender and living standards affect a person's educational attainment? How can we expect educational attainment to change in the next few years due to gender and living standards?

Proposed approaches

1. For our first question we must explore the relationship between educational attainment and wealth. To explore this relationship it will be useful to first explore the distribution of each variable on its own. We would explore educational attainment by plotting against the observed proportion to see the proportion of the sample that attained the education at each grade level across all countries combined. We can then build a seperate scatterplot of our variables prop vs. gpdpc2015 where we can encode the points within the scatterplot to represent another variable such as the countries to give us more information about our data.

2. For our second question we are looking into how gender and living standards affect educational attainment. For this question we would explore the relationship of these variables by plotting educational attainment vs year. We would encode each plot to show the proportion of attained education with a different line for each location (whether rural or urban) and be faceted by sex to allow us to have a more clear visual of any significant trends that are attributed to either sex or location. Given that earlier in this report we could see that there is a notable relationship between these variables, our next step would be to fit our data to a multiple linear regression model given that we have data for rural and urban locations seperately. After fitting our data would go back to creating a line plot similar to the one previosuly done, plotting prop vs year and mapping location to the color aesthetic. Lastly we would make use of the data given by our fitted regression model to predict how we would expect educational attainment to change in the next few years with our variables of interest.

Submission Checklist

1. Save file to confirm all changes are on disk
2. Run *Kernel* > *Restart & Run All* to execute all code from top to bottom
3. Save file again to write any new output to disk
4. Generate PDF and submit to Gradescope