R exercise 2: Data Preprocessing
INFO 523
Group 1
Matt Miller
Kai Blumberg

1. What attributes are there in your data set?

The Gapminder dataset https://www.gapminder.org/data/ has over 500 attributes for various economic, societal, health, and other indicators for countries across different time frames. However, each attribute is in it's own CSV file, with different countries, years, and numeric representations (percents, in millions, in billions, etc.), so we focused on merging attributes together and handling disparate information. For this exercise, we started with 4 attributes: % adults with HIV (ages 15-49), % of the population that is female ages 40-59, gross domestic product purchasing power parity (GDP PPP), and % underweight children. We will be downloading and adding more attributes to our dataset as the project progresses and we think of interesting relationships to explore.

2. Do you have highly correlated attributes? How did you find out about the correlations or lack of correlations?

All four of our attributes are numeric, so we calculated the covariance and correlation coefficient to determine whether our attributes are correlated. We found:

```
> hiv_vs_weight_cov
[1] -3.150009
> hiv_vs_female_cov
[1] -5.611719
> hiv_vs_gdp_cov
[1] -339348943427
> weight_vs_gdp_cov
[1] -2.340834e+12
> weight_vs_female_cov
[1] -24.76633
> gdp_vs_female_cov
[1] 1.052022e+12
> hiv_vs_weight_corr
[1] -0.0543583
> hiv_vs_female_corr
[1] -0.2939716
> hiv_vs_gdp_corr
[1] -0.1367245
> weight_vs_gdp_corr
```

```
[1] -0.1247399
> weight_vs_female_corr
[1] -0.42669
> gdp_vs_female_corr
[1] 0.2456312
```

- % HIV Adults is very weakly negatively correlated with % Underweight Children
- % HIV vs % middle-aged females, % HIV vs GDP-PPP, % Underweight vs GDP-PPP, and % Underweight vs % middle-aged females are all moderately negatively correlated, with % Underweight and % middle-aged females having the strongest correlation. This means that a higher percentage of middle-aged females is associated with a lower percentage of underweight children. We could look into these attributes relationships with childbirth mortality rates, or female income.
- GDP-PPP and % middle-aged females are the only positively correlated attributes.

Drilling down to further investigate the hiv vs underweight children correlation, which in the aggregate had a near 0 correlation, when we group by countries we find very variable correlations:

Country	**hiv_vs_weight_pearson_corr**
Angola	-1.00000000
Benin	0.18591675
Burkina Faso	-0.98626197
Cambodia	0.98980812
Cameroon	-0.57648582
Chad	-0.13719939
Colombia	-0.83576611
Cote d'Ivoire	0.56685043
El Salvador	-1.00000000
Equatorial Guinea	-1.00000000
Gambia	1.00000000
Ghana	-0.53181570
Guatemala	-0.93532167
Guinea	-1.00000000
Guinea-Bissau	-1.00000000
Guyana	1.00000000
Haiti	0.53776683
Honduras	0.98856972
India	1.00000000
Indonesia	-0.58615218
Jamaica	0.26124530
Kazakhstan	-0.13802653

Kenya	0.09469825
Lesotho	-1.00000000
Liberia	1.00000000
Malawi	-0.14298339
Mali	0.89483274
Mauritania	-1.00000000
Morocco	1.00000000
Mozambique	-1.00000000
Myanmar	1.00000000
Namibia	-0.67942290
Nepal	-0.05591918
Nicaragua	-0.91969828
Niger	0.41832768
Nigeria	-1.00000000
Pakistan	-1.00000000
Peru	0.34398138
Rwanda	0.99985630
Senegal	-0.90828937
Sierra Leone	1.00000000
Sudan	-1.00000000
Swaziland	-1.00000000
Tanzania	0.77273603
Togo	1.00000000
Uganda	0.95232528
Uruguay	1.00000000
Uzbekistan	-0.69161285
Vietnam	-0.95006549
Zambia	0.43626598
Zimbabwe	-0.97899234

Where we note that many countries have pearson correlation values near -1 such as Zimbabwe or 1 such as Uruguay. These data can be explained by in some cases such as for Morocco, there only being 2 data points with both HIV and underweight children, hence the great variability in slopes being positive or negative.

3. Do you have numerical attributes that you might want to discretize? Try at least two methods and compare the differences

Attempting to bin by groups of years to re-examine the correlation between hiv and gdp we get the following which is a bit hard to interpret

Bin (year)	**hiv_vs_gdp_pearson_corr**
(1990,1994]	-0.3077308
(1994,1997]	-0.1009098
(1997,2001]	-0.1975727
(2001,2004]	-0.1147347
(2004,2008]	-0.1875082

Attempting to discretize the hiv infection rate to do the same correlation we get:

Bin (hiv)	**hiv_vs_gdp_pearson_corr**
(0.0341,8.71]	-0.1206205
(8.71,17.4]	-0.3344022
(17.4,26]	0.3382484

Which is actually interesting because the countries with the highest HIV infection rates have a slight positive correlation between HIV infection rate and GDP, whereas the countries in the lower HIV infection rate bins have a slightly negative correlation signal. Essentially for certain countries with **really** high HIV rates, the HIV rate correlates slightly to GDP, but for lower or average HIV infection rates it doesn't correlate to GDP.

4. If you have categorical attributes, use the concept hierarchy generation heuristics (based on attribute value counts) suggested in the textbook to produce some concept hierarchies. How well is this approach work for your attributes

We do not have any categorical attributes at the moment, however, as we are dealing with country data, it would be possible to group countries with concept hierarchies. We could group countries by geographic region: Europe, Asia, Africa, North America, Middle East etc. Alternatively there may be some existing categories which we could leverage for example development status, i.e. developed nations, developing nations, etc.