1. What's the name of your final project?

Research question: What are some factors that correlate to a country's happiness ranking?

I collected data from two sources:
1) One openly available dataset from 2015 called the "World Happiness Report" from Kaggle (https://www.kaggle.com/datasets/unsdsn/world-happiness) and 2) using country demographic statistics from a country API (https://api-ninjas.com/api/country).

My project aims to identify correlations and trends between a country's demographic statistics (i.e. factors like infant mortality rates, internet users, geographical location/region, and urban population) and its happiness ranking.The analysis method I used include correlation analysis using NumPy and Pandas, and creating visualization, like scatterplots, correlation heatmaps, bar charts, and maps that show the relationship between a country's happiness ranking and specific attributes.

2. How to run your code?

My project can be run as a Jupyter Notebook (main.ipynb) by selecting "Run All Cells". There is nothing specific to do as there is no need to run multiple files or steps. All of the code is in the main.ipynb file, and dependencies are listed in requirements.txt.

Github Repo: https://github.com/talajune/DSCI510_FinalProject

Dependencies

requests==2.27.1

requests-file==1.5.1

pandas==1.4.2

numpy==1.21.5

matplotlib==3.5.1

plotly==5.6.0

seaborn==0.11.2

Installation

How to install the requirements necessary to run my project: You can use the requirements.txt that is in my project folder.

pip install -r requirements.txt

3. What data did you collect? How did you collect them? How many data samples did you collect?

I collected data from two sources:
1) One openly available dataset from 2015 called the "World Happiness Report" from Kaggle (https://www.kaggle.com/datasets/unsdsn/world-happiness). This was very straightforward, as I downloaded the CSV from Kaggle and simply read it into my Jupyter notebook using Pandas. Below are the first few rows of the happiness_ranking_data dataframe.

Overall rank	Country or region	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
1	Finland	7.769	1.340	1.587	0.986	0.596	0.153	0.393
2	Denmark	7.600	1.383	1.573	0.996	0.592	0.252	0.410
3	Norway	7.554	1.488	1.582	1.028	0.603	0.271	0.341
4	Iceland	7.494	1.380	1.624	1.026	0.591	0.354	0.118
5	Netherlands	7.488	1.396	1.522	0.999	0.557	0.322	0.298

2) Country demographic statistics from an API Ninjas country API (https://api-ninjas.com/api/country).

Collecting data from the API was a little more challenging as there was a limit on the number of results it could return at once, and the country name had to be fed in as a parameter, but it only allows for one country name at a time, per request. So, I created a list of the country names that are included in the World Happiness Report dataset, and then created a for loop that would go through each country in the list and get the demographic data for that country and append it to a new list, making a list of list of dictionaries. Then, I converted the list into a new dataframe using Pandas and chain.from_iterable(). From there, I dropped columns I wouldn't use in my analysis, merged the two dataframes on the "name"/"country or region" key using Pandas.merge, and changed the column order.

After data pre-processing the new dataframe (ie. dropping rows with NaN values, dropping columns I would not use in my analysis), I was left with a final dataset of 116 rows (data samples/instances) and 21 columns (features).

Below are the first few instances of the api dataframe, with only some columns visible.

urban_population	secondary_school_enrollment_male	name	region	pop_density	internet_users	gdp_per_capita	fertility	refugees
85.4	146.8	Finland	Northern Europe	18.2	88.9	50135.7	1.5	35.0
88.0	128.4	Denmark	Northern Europe	136.5	97.3	61833.7	1.8	47.4
82.6	119.4	Norway	Northern Europe	14.8	96.5	81335.7	1.7	59.9
93.9	118.1	Iceland	Northern Europe	3.4	99.0	76867.3	1.8	1.2
91.9	134.7	Netherlands	Western Europe	508.2	94.7	53583.1	1.7	109.1

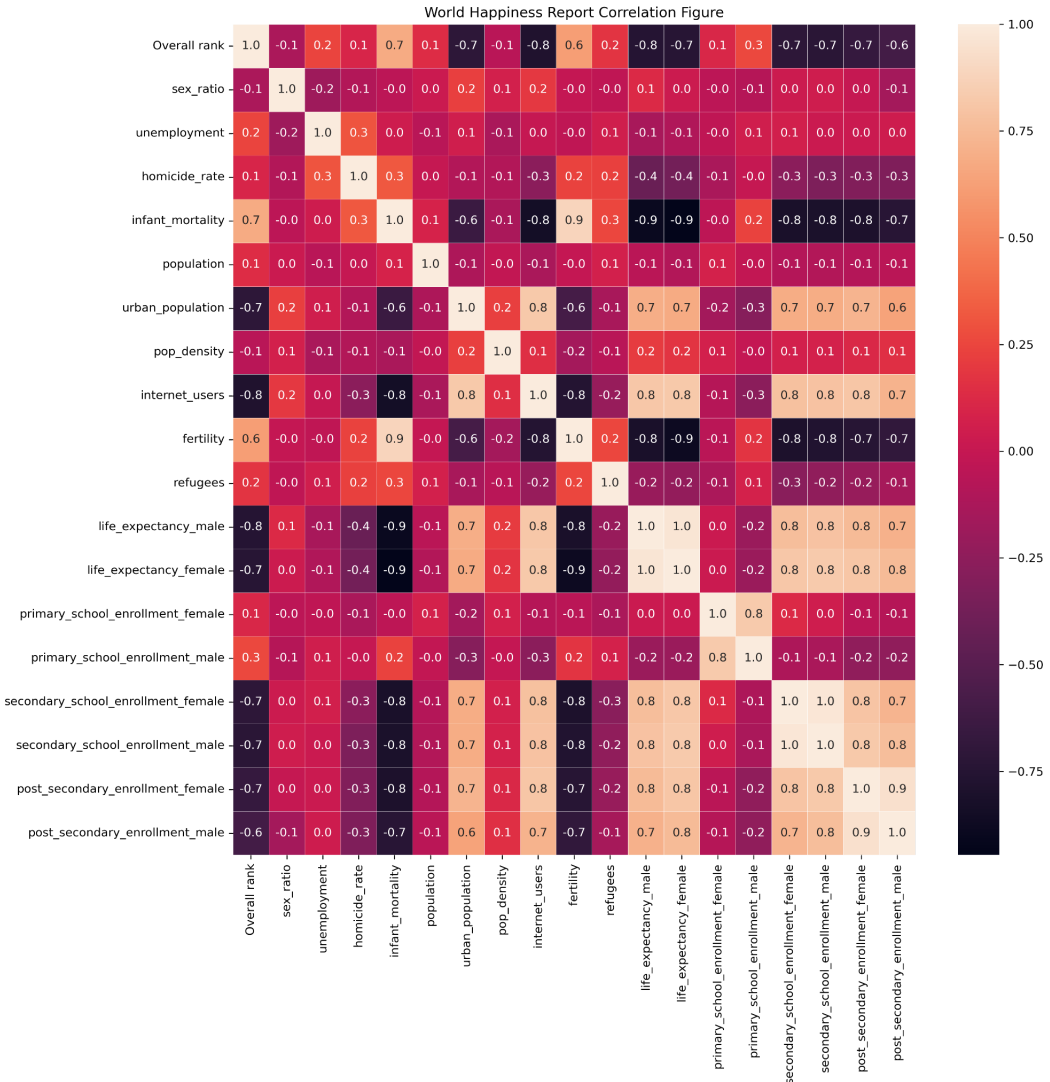
b. Describe what has been changed from your original plan, what challenges you encountered or resolved.

Initially, I was interested in looking at a select number of countries, specifically, the top 30 “happiest” countries, to try and see what feature values and demographic trends they had in common. However, I realized later that it would be best to include all of the countries I had data for, so that I would be able to draw conclusions and perform analysis on a wide range of happiness rankings and demographic statistics. I thought that including more data instances, especially from countries with lower happiness rankings, would help my visualizations be more representative of trends as happiness ranking decreases.
I ran into some issues along the way with how to get data for 100+ countries from the API since it only gets data for one specific country at a time, but the TAs helped me work through that by creating a for-loop. There were also some other challenges I ran into, such as some of my visualizations not showing up when running the entire notebook, but a few StackOverflow searches helped me fix that. As I did more of a deep dive in the analysis, visualizations, and got to know my dataset better, I made some changes to my initial plan and did not include logistic regression models or ML models and focused more on simple, correlation visualizations.

4. What kind of analyses or visualizations did you do?

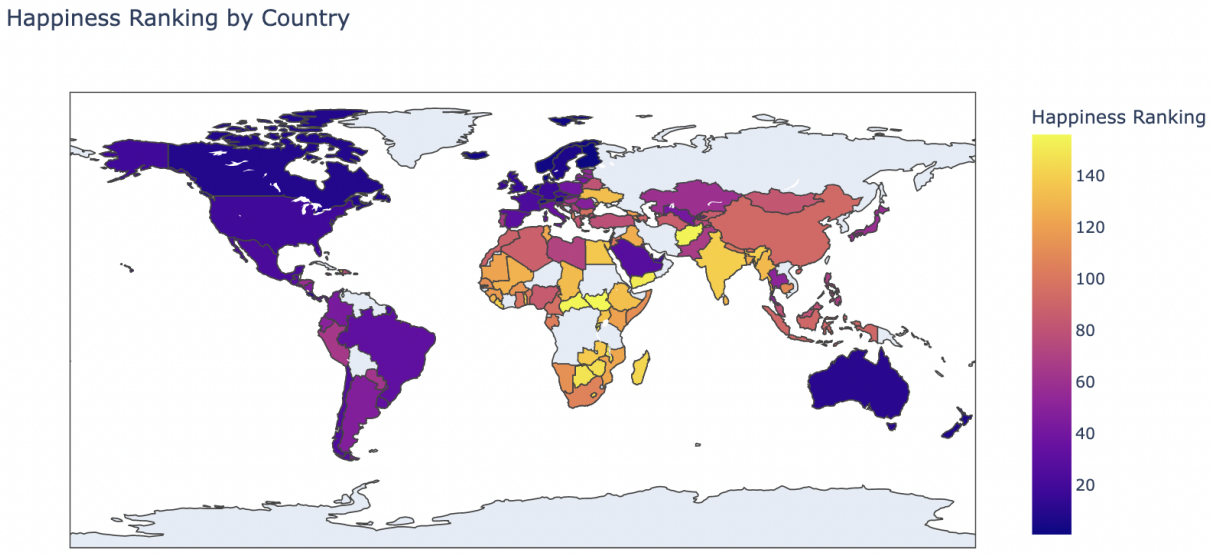
One thing that is important to note for this project and remember throughout the analysis is that due to the ranking of the countries in the dataset, the “happiest” countries are ranked higher on the list, meaning that countries with a rank of 1 through 10 are the top 10 happiest countries, and countries with a rank of 140 to 150 are the “unhappiest” countries, ranked “lower”.

Initial Analysis Using a Correlation Heatmap



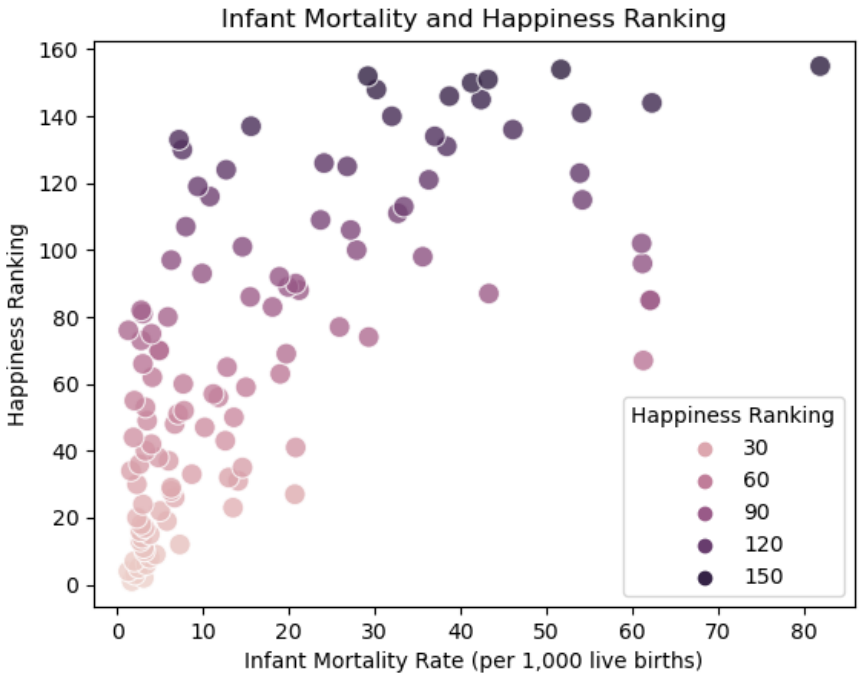
I wanted to use this correlation heatmap to better understand the correlation relationship between different features in my dataset. After dropping columns I felt were irrelevant to my project, I created this correlation heatmap as a first step to get a better understanding of the data. features. My main takeaways from this correlation analysis is that it looks like the features that are most correlated to “Overall rank” (Happiness rank), whether positively correlated or negatively correlated, include urban_population, internet_users, life_expectancy_female, life_expectancy_male, infant_mortality, and the different education enrollment features (ex. primary_enrollment_female). Since the happiness ranking is conceptually inverse (as happiness rank increases, happiness actually decreases), then the factors that are strongly negatively correlated, like internet_users (-0.8), actually shows us that as the internet users in a country increase, the country’s happiness rank decreases (meaning that their “happiness” actually increases – they are a happier nation).

Map Visualization of Happiness Ranking



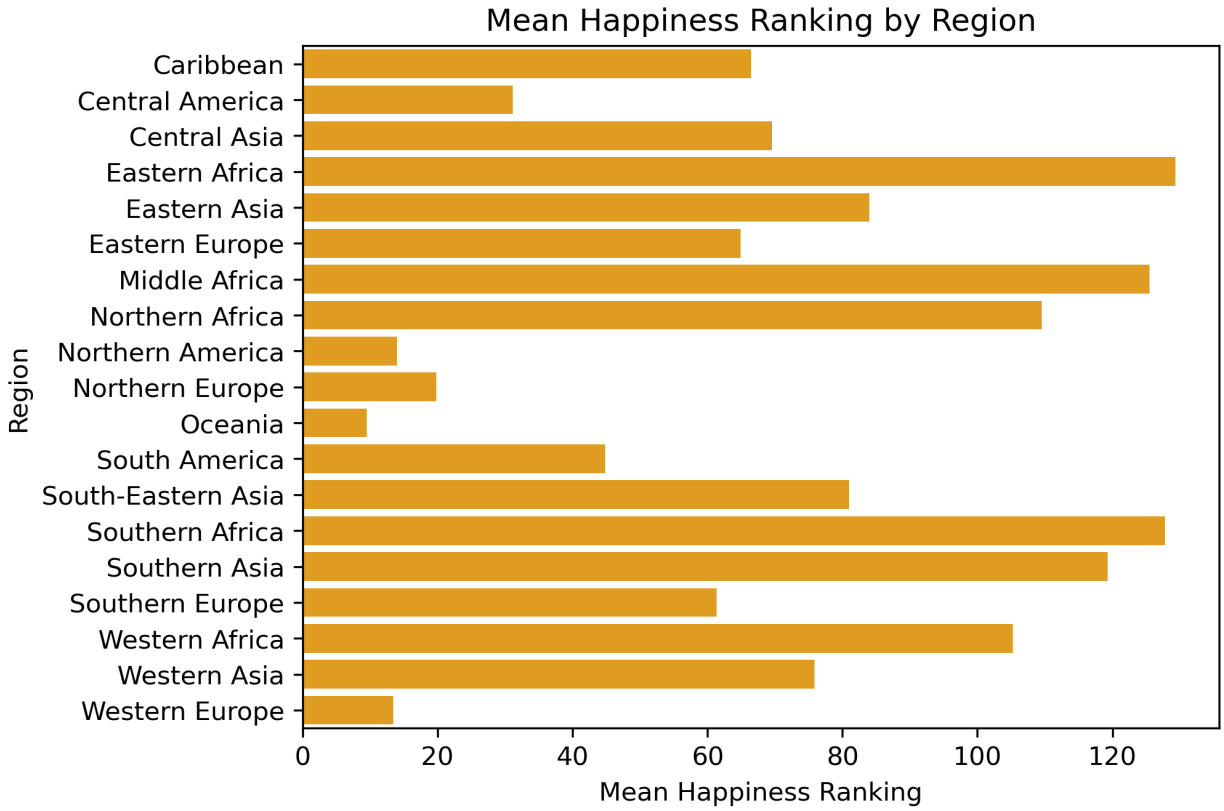
This was more of a fun visualization, just for me to be able to visualize how the happiest countries are dispersed across the world. There is data missing for some countries that were not included in the original World Happiness Ranking dataset from Kaggle, which I assume is for various reasons, including a lack of reliable data, a lack of reported data, or even political reasons (for example, countries with land disputes or civil wars). These omitted countries include Iran, Russia, Greenland, and Venezuela. It is interesting to see how the rankings vary by continent, with countries in North America and Northern and Western Europe reporting as “happier” than countries in Africa or Asia. One country that stands out to me is Saudi Arabia, I wasn’t expecting their Happiness Ranking to be so low on the ranking list, we can see they are one of the only middle eastern countries to have such a low happiness ranking (they are happier than most!)

Infant Mortality and Happiness Ranking Scatterplot



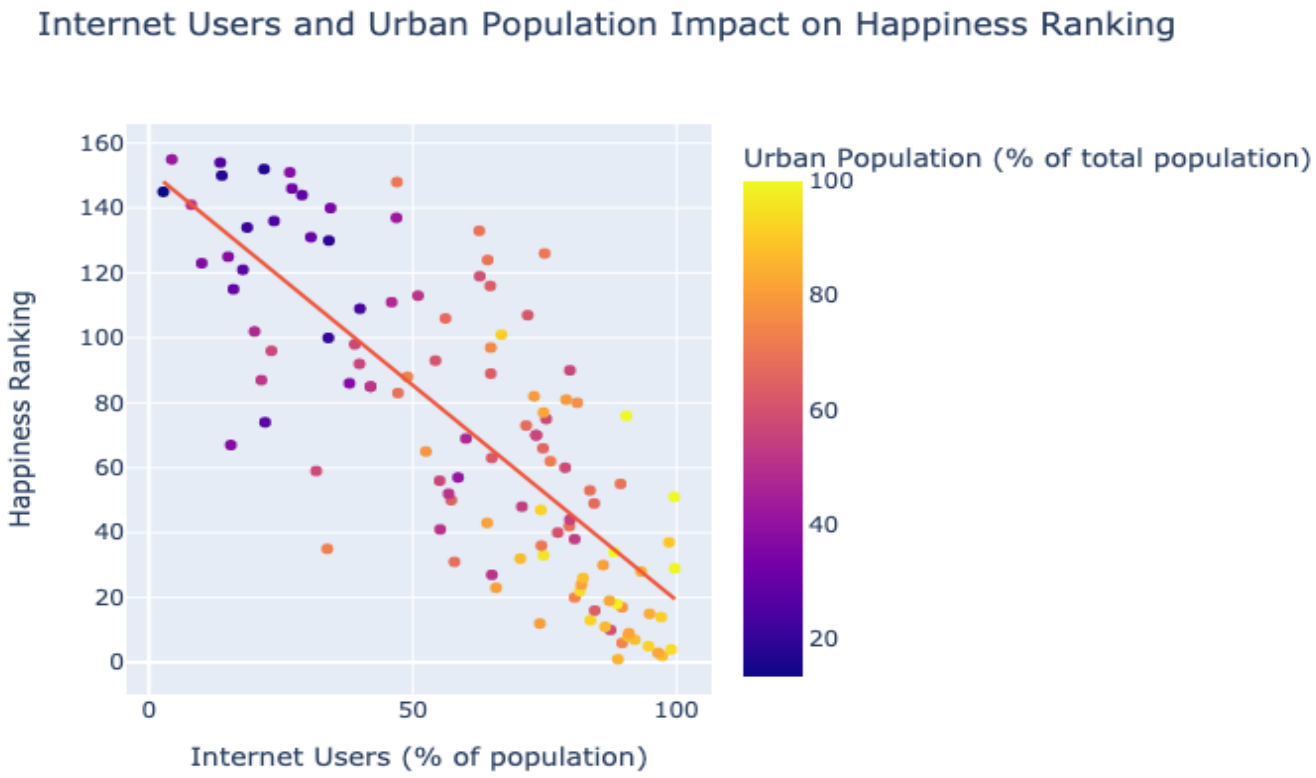
As seen in the correlation graph above, infant mortality has a high correlation with happiness ranking (0.7). I wanted to visualize this to get a better understanding of the relationship, and we can see that as infant mortality rates increase, the happiness ranking consistently increases (countries get less happy). Almost all of the countries that are happiness ranked between 1 and 60 have infant mortality rates less than 15, with only a few visible countries on the graph having low infant mortality rates and being ranked very high on the happiness index.

Mean Happiness Ranking by Region Bar Chart



This bar graph shows the mean happiness ranking for each region of the world. We can see that Oceania, Northern America, Northern Europe, and Western Europe have the lowest average happiness ranking, making them the “happiest regions”. Regions like Eastern Africa, Middle Africa, and Southern Africa have the highest average happiness ranking, making them the least happy regions. I computed the mean happiness ranking by grouping the dataset by their respective regions and finding the mean for each region’s overall happiness rank, and making it into a new dataframe.

Internet Users and Urban Population, Happiness Ranking Interactive Scatterplot



This visualization is interactive (hover over points for specific data points like specific country name) and gives a glimpse into the relationship between percentage of population that are internet users, the country's happiness ranking, and it is colored by the country's percentage of total population that live in urban areas. The relationship presented by the graph is very clear: as internet users in a country increases, the country's happiness ranking gets lower (happier and closer to 1). Additionally, it looks like the relationship is the same for the urban population, as the percentage of the total population that lives in urban areas increases, we can see a general trend that the country's happiness ranking decreases (they are happier countries). We can see that the happiest countries also have the highest percentage of internet users and highest urban population.

Conclusion

As my visualizations and analysis have shown, some of the strongest correlated features in my dataset related to happiness ranking include the percentage of a population that are internet users, percentage of population that live in urban areas, low infant mortality rates, education rates, and life expectancy for men and women. These all are reasonable indicators in my opinion, and they make logical sense. For example, countries with low infant mortality rates probably afford their citizens good access to healthcare, and better immunizations and disease prevention. Countries with more internet users can lead to its citizens using the internet to educate themselves on various topics, find work, stay connected with friends and family, and share ideas, information, and knowledge. As I expected before the project, I thought that the happiest countries would be in North America and West and Northern Europe, which was supported by my map visualization and bar chart. Finally, I was surprised to see in the correlation heatmap that homicide rate was not strongly correlated to happiness ranking, since I would assume countries with more crime would lead to its citizens to feel more unsafe, uncomfortable, and stressed rather than leading to a happier nation.

Future Work

a. Given more time, what direction would you take to improve your project?

If I had more time, I would probably import and join a separate dataset that had features that I didn't have included in the data I collected from the API, for example, I would like to find some features that are better predictors or have higher correlations to the Happiness Rankings for each country. Perhaps I would try to incorporate data features in categories such as freedom of speech, access to healthcare statistics, police brutality data, or political data. In short, I would expand the features I used by including more data sources and try to find more correlations between different features and happiness ranking.