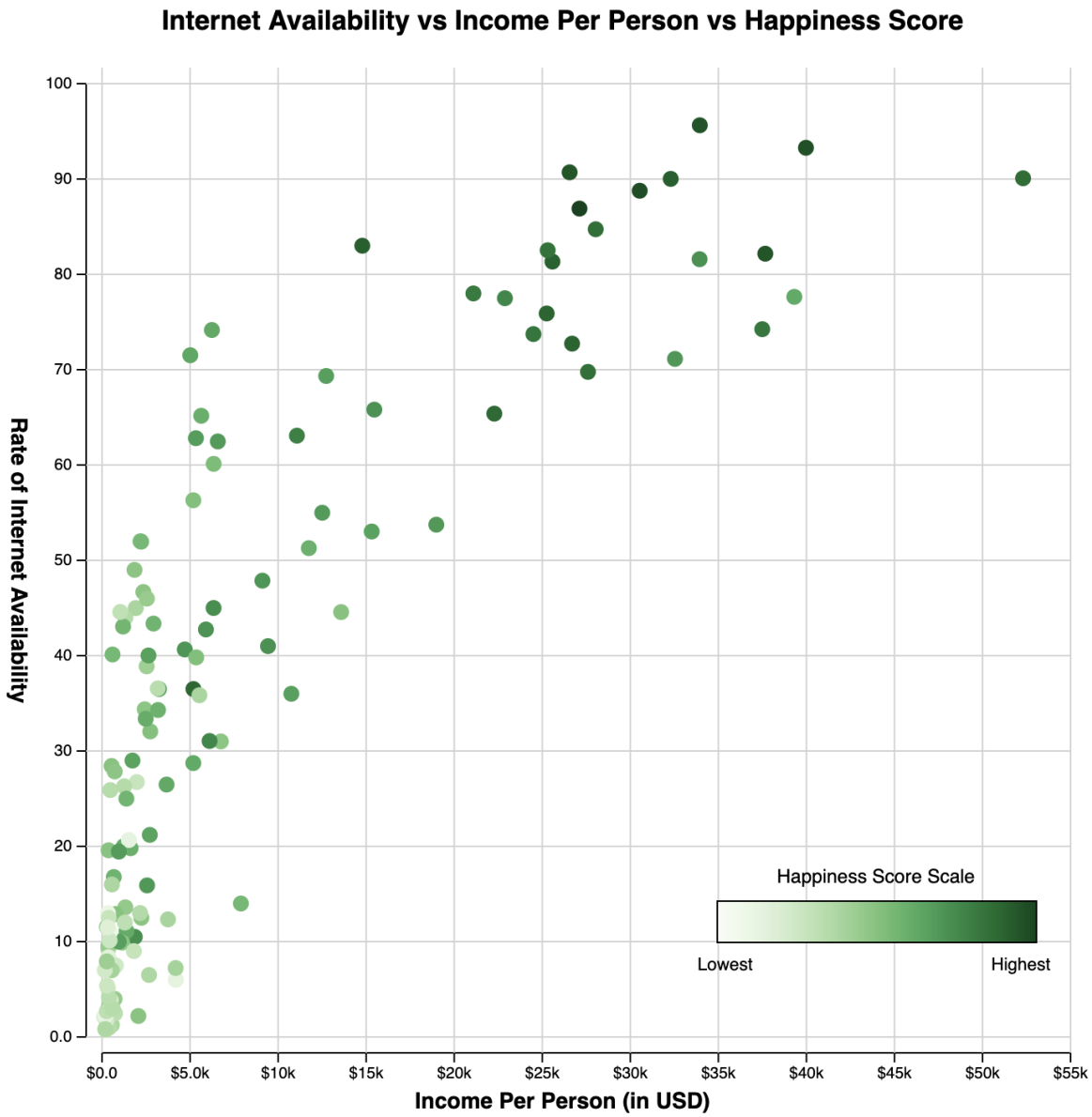
Group members: Adam Kadhim , Nada Attia , Jacob Ball , Akshay Yadava
CS 3300 Project 1 Final Report



Description of the Data

We had taken both data sets from kaggle.com, a website for publicly accessible data sets. From the first data set, we plotted the internet use rate and income per person for each country; from the second data set, we used the happiness score from each of the corresponding countries we had looked at from the first data set and represent the relative happiness using shades of green for each data point.

In order to pre-process the data, we removed all instances of *‘’* and *undefined* in order to not represent any incomplete data points. Additionally, we only represented countries that were present in both data sets; this was done from a usability perspective as coming up with ways to represent points that were inconsistent in information with the rest would confuse the viewer. We only did this for data points of interest and did not examine if a specific field we did not use of the csv was incomplete. Additionally, we accomplished the pre-processing by examining each data point before using them and filtering based on the previously specified criteria.

We recognized since we only needed to look at each point once for the plotting of the data that we could just run parallel loops to match both data sets. We did this by: for each country in the internet dataset, we tried to find the country in the happiness data set and then plotted the information from both datasets and moved on to the next country. As stated previously, if we do not find the country in both datasets, we do not use the data point.

Overview of Design Rationale

1. Circles (Scatterplot)

We decided to create a scatterplot of the data we found. Since we wanted to visualize two separate datasets (one for the happiness scores per country, and one that includes internet usage per country as well as income per person per country), we decided to group all of this data by countries since country name was a common attribute in both datasets. Each circle in this graph represents a country and the location of the circle on the chart conveys the rate of internet availability of the country as well as its income per person.

2. Color of the circles

We decided to vary the luminosity of the circles to visualize the happiness score of each country. The happiness scores are on a scale of 1 to 10. Dark green represents a happiness score of 10 whereas light green represents a happiness score of 1. We chose a sequential scale because we wanted to map scores to a color to differentiate between low and high scores. We chose the color green specifically because we think it fits the data attribute that we wanted to represent (happiness).

3. Scale and Gridlines

The y axis represents the rate of internet usage so we decided to use a 0 to 100 scale for the y axis and add grid lines at intervals of 10. We thought this was the most appropriate scale for percentages. The x axis represents income per person so we decided to use a 0 to 55k scale with grid lines at 5k intervals. The reason we chose those specific scale intervals instead of the x extent and y extent was because we didn’t want any data points to be on the very edge of

the graph so we chose the max value on the x and y axes to be close to the max value of the x and y data (but not exactly the max value of the data).

4. Title, Labels and Legend

We added a title to describe all the data that the chart aims to visualize. We added labels on the x and y axis to show what data each of the axes represent. We chose bold fonts for all of that text to emphasize it since it is hard to make sense of the data without and y labels as well as a title. We also added a legend to show what the different luminosities of the circles represent. This reinforces the association of dark green colors with high happiness scores and light green colors with low happiness scores, and clears up any confusions the viewer might have.

The Story

Fundamentally, we expect all viewers to first have their attention drawn to the title of the graph, continuing a tradition of literary etiquette thousands of years old. Then, falling upon the viewer, they are confronted with the influx of meanings and a near infinite set of experiences from their lifetimes shaping the words “income”, “happiness”, and “internet”, all of which are subject signifiers for a significance only known to the individual. From which they will then see each individual point on the graph and notice the variation in color and see that the color increases in saturation from left to right and begin drawing conclusions about happier people making more money which likely fits their prototypical thoughts.

They will then examine the relationship between happiness and rate and internet availability. It appears as if countries with a higher rate of internet availability also have higher rates of happiness. Although we do not know for sure if the increase in happiness is a direct result of the country's internet availability it is still an interesting revelation. There is also a clear correlation between increased internet availability among countries with a higher average income per person. This observation confirms our initial hypothesis that wealthier nations have better access to the internet.

It is important to note that people living in various countries cannot control the socioeconomic factors that are prevalent in their country. People are born into their home country at random and must live with the realities of their country. Some people are born into relatively more impoverished nations at random and thus may face more obstacles in obtaining wealth and reaching financial prosperity. This visualization shows that there is a seemingly positive correlation between average income per person and average happiness rating of a country. This visualization tries to illustrate to the reader that though there is a seemingly positive correlation between these two factors, one’s happiness level should be derived intrinsically free from external influencers as these factors tend to vary heavily on a country to country basis.

The data in no way provides a complete picture into the full correlation between happiness level, internet usage, and average income per person on a country-wide basis. Yet,

this visualization provides some insight into the perceived relationship between these variables and how different variables are related positively or negatively with one another.

Team Contributions:

Nada:

- Set up the x and y scale.
- Drew the gridlines/axes.
- Added margins to the graph to make space for the labels.
- Added x and y labels, and the title.
- Worked on adding a color scale legend.

Adam:

- Found data sets online and created ideas for stories.
- Created initial rough draft visualizations of different data sets to see what presented a story we might want to tell.
- Filtered both data sets to get rid of empty entries.
- Cleaned code to be more efficient and refined.
- Made Legend for chart.
- Graphed points and set up loop structures for cross referencing datasets.

Akshay:

- Constructed initial graph of income per person vs. rate of internet usage
- Incorporated sequential legend to interpret the various happiness scores of the data points
- Attended in class critique and updated legend corresponding to peer feedback
- Helped with the Story portion of the final report

Jacob:

- Limited our choices for final datasets
- Filtered datasets to remove countries that were not represented across the multiple datasets
- Incorporated the color scale to represent the happiness score
- Attended in class critique and made corresponding adjustments; updating the legend to account for saturation.

After we found the datasets that we wanted to use, we overall spent around 10 hours developing this data visualization. The part of this project that took the most time was adjusting the appearance of the graph and the legend as well as preprocessing the data. This was our first time making a legend of this sort so figuring out how to do it and debugging it took a fair amount of time. We also spent time adjusting the x and y scale to make sure that the data wasn't too crammed but also not too close to the edge of the chart area. Since we were combining two different datasets to make one graph, we had to preprocess the data. Looking through the data and understanding how to combine them together also took a while.