DECO3100
ASSIGNMENT 2

Exploratory Data Analysis

Andrew Xing
Monica Femenias
Hyo Kyung Kim

Table of Contents

1. Introduction

Suicide rates in South Korea have always been notoriously high, being the most common cause of death in people aged 10 to 39, and second most common among adults aged 40 to 59 (Lee et al., 2018). Being ranked as having the 10th highest suicide rate in the world by WHO, much of the issue can be traced back to its toxic social climate. According to (Singh, 2017), it is the only OECD (The Organisation for Economic Co-operation and Development) country in which their suicides rates have increased since the 1990s. For the youth, bullying runs rampant in schools, and intense pressure is placed on a student's academic success, with results being publicly posted on school boards, while adults face issues such as income inequality and work overloading (Goh, 2019). Exploring the suicide rates for each age group and gender may shed light on how each generation is affected

2. Hypothesis

We propose that suicide is a more prevalent issue among the youth in South Korea.

3. Data Description

The original dataset was retrieved from Kaggle, from a user who compiled it from four other datasets linked by a time and place. Its purpose is to find trends correlated to increased suicide rates amongst different countries across the globe; providing information starting from 1985 until 2015. Due to the sheer volume of information, we decided to condense the data and focus solely on South Korea, as it is a country is well known for its high suicide rate, which is a prevalent issue there. We also thought that it would be more interesting to do this, and that we would be able to visually represent this data more effectively. We've also decided to discard the 'generation' column, as we felt that it was unnecessary and analysing the age groups alone would give us more information.

Columns

Year: The year in which the data was recorded

Sex: The gender; whether they are male or female

Age: The age group of the recorded data

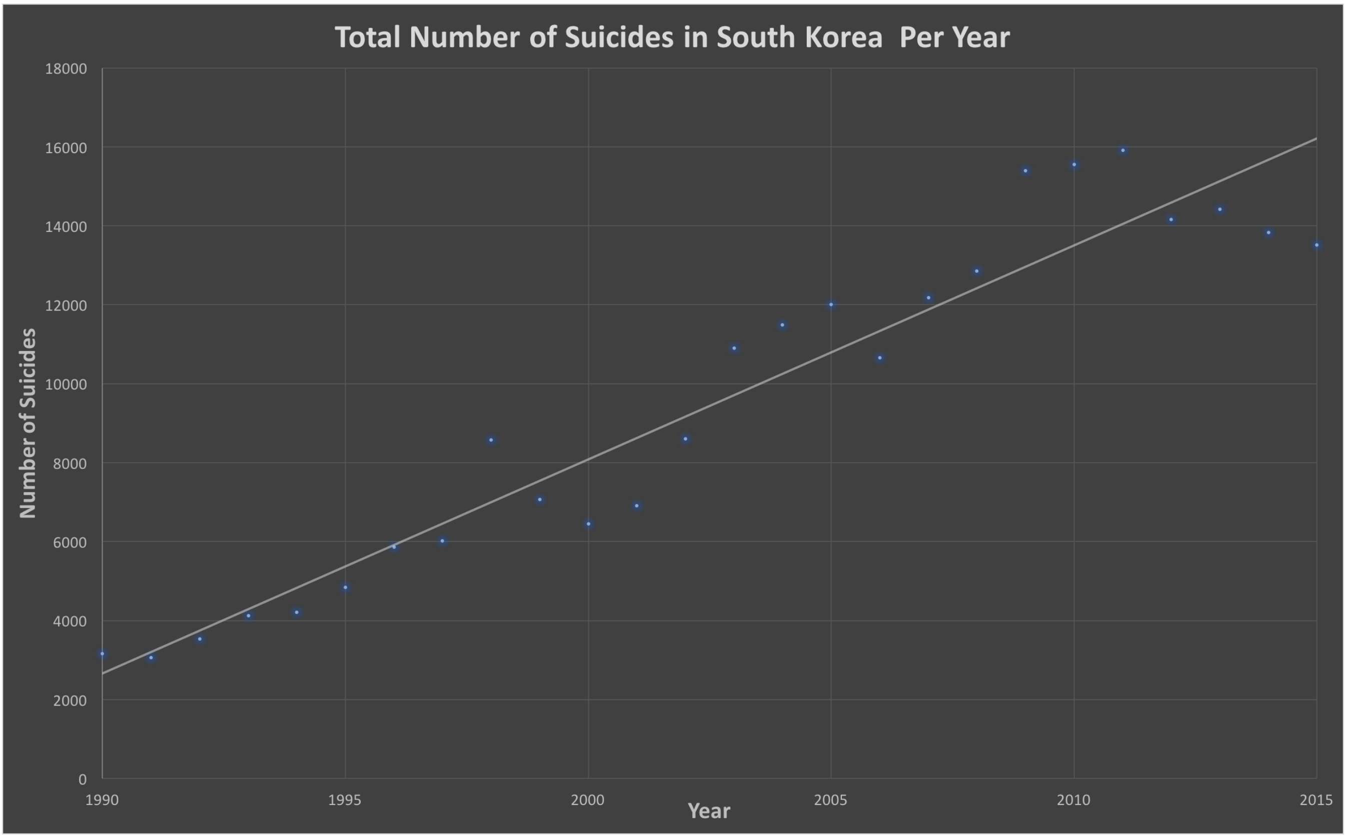
Suicide no.: The number of suicides for that age group

Suicides/100k pop: The number of suicides for that age group per 100k people

GDP_per_year: The Gross Domestic Product (GDP) per year

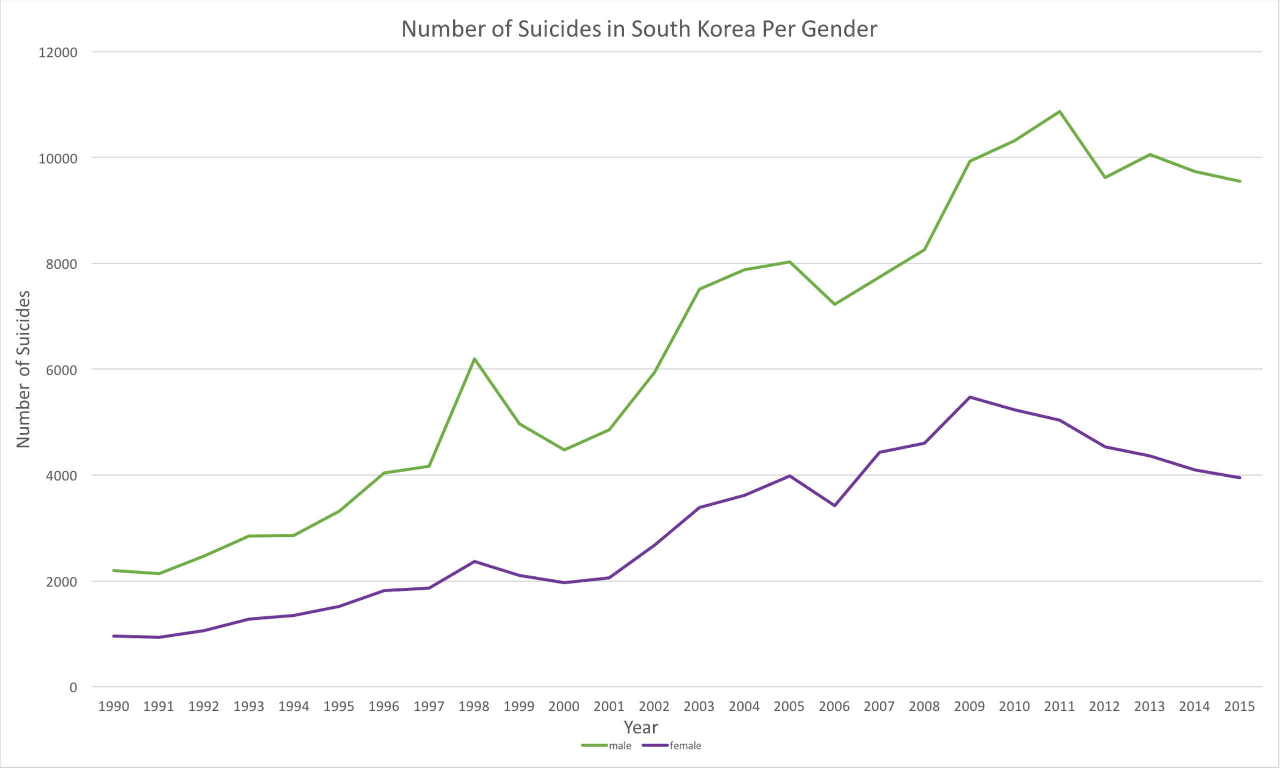
4. Exploratory Data Analysis

Scatter plot



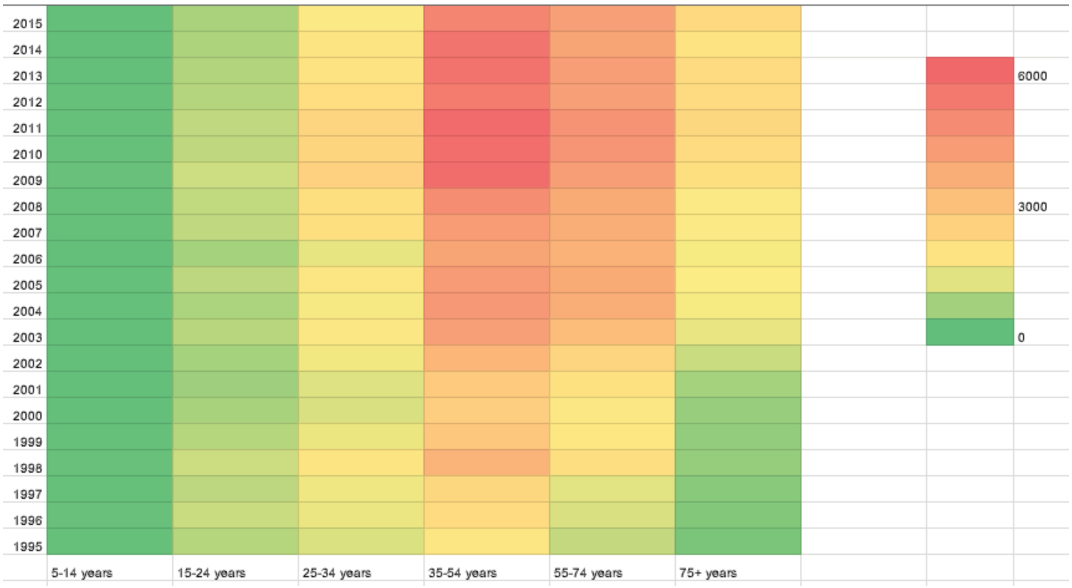
The first graph we decided to create was a scatter plot using excel. Here we visually represented the overall amount of suicides per year, starting from 1990 until 2015. By including a trendline, we were able to see that the number of deaths is definitely increasing. The overall number begins at around 3000 in 1990 and soars up to as high as around 16000 in 2011. We concluded it was an effective way to view the entire spectrum and attain a relatively basic understanding of the overall scheme as an initial step. We observed the rate of deaths beginning to increase at a steady rate throughout 1990 until around 1997. Following this, the rise became very unpredictable, with the amount of suicides per year quickly increasing and decreasing at an unsteady rate. There also appeared to be an increase of around 2000 additional deaths every five years. From this visualisation, we could go further into depth in our next explorations.

Gender line



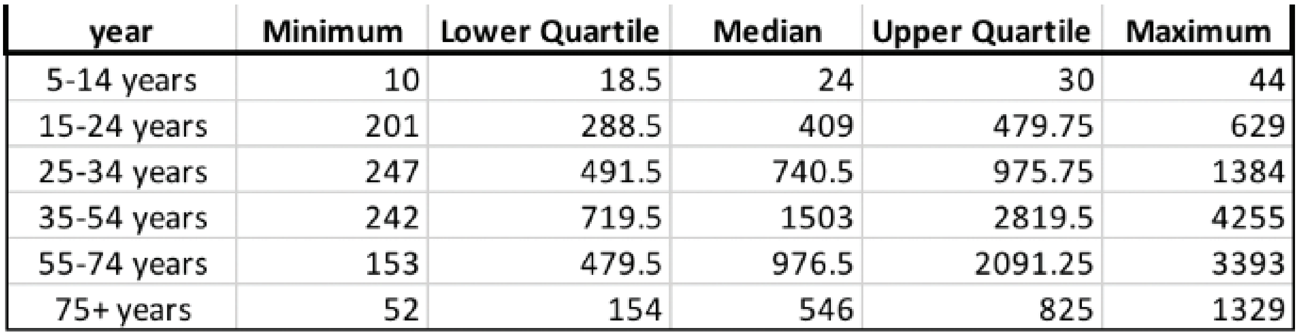
As a further breakdown of the scatter plot, we had created a line graph that depicted both male and female rates throughout the years. This was achieved by calculating the total number of females and males for each year using the =sum() function on excel. It depicts the trends between each gender, with the purple line representing the female rate, and green being the male rate. By analysing the graph, we observed that the number of males who had passed was almost double the amount of females. Additionally, the female line was shown to be a lot more steady, with the male line constantly jumping up and down. Both gender rates spiked during 1998 and 2009, but quickly decreased again the next year. The rate of deaths for both genders seemed to be decreasing since around 2013, but had both been predominantly rising throughout the years prior. Both the female and male rates had also generally been on the same path, rising and falling during the same years. By gender, the male suicide rate had increased significantly from that of women, indicating that men are more vulnerable to economic crises than women.

Heatmap



The heatmap allowed us to digest and understand the information more easily, expressing differences in color, which were used to visually compare the number of suicides corresponding to six ages (5-14, 15-24, 35-54, 55-74, 75+ years) from 1995 to 2015. From the data, several trends were noted. The red series of colors appeared most prominently within the 35-54 age column, followed by the 55-74 age column. On the other side, the predominantly green 5-14 age column indicated its extremely low suicide numbers. Furthermore, the 75+ age column featured a gradually rising trend, whilst the 15-24 age and 25-34 age columns remained at a relatively similar number.

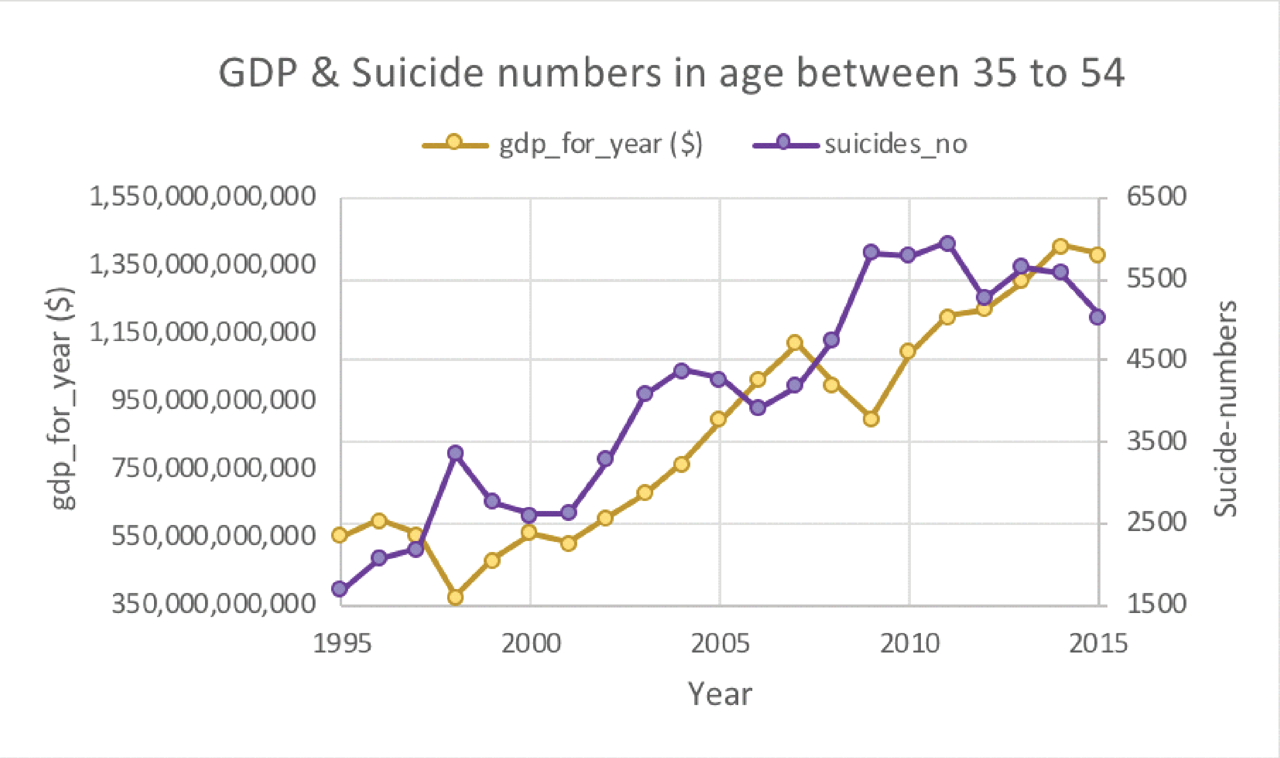
Five-Number Summary



year	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
5-14 years	10	18.5	24	30	44
15-24 years	201	288.5	409	479.75	629
25-34 years	247	491.5	740.5	975.75	1384
35-54 years	242	719.5	1503	2819.5	4255
55-74 years	153	479.5	976.5	2091.25	3393
75+ years	52	154	546	825	1329

We used the five-number summary as a way to summarize the dataset using the following five values: minimum, first quartile, median, third quartile, maximum. Unlike the heatmap, in order to access the data more mathematically and conduct accurate and in-depth analysis, we created summary using suicide case data corresponding to each age category from 1995-2015. The above image was the result of five number summaries, and because the ranges held by each category data were so widely different, it was not appropriate to make a box and whisker plot.

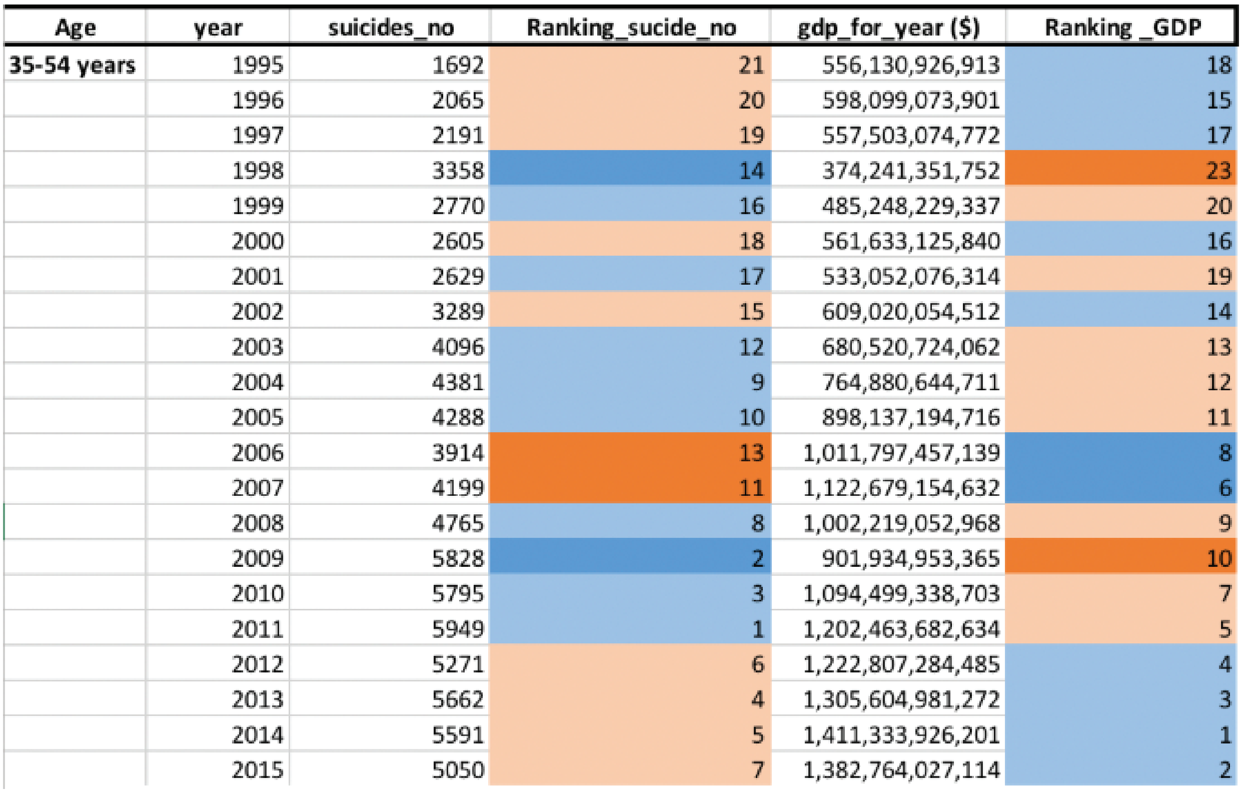
GDP Scatter plot



According to our heat map, the 35-54 age group showed the highest rates amongst 6 different ages categories. Based on this finding, we limited the age range to 35-54 years and compared their numbers against the GDP to look for any potential trends.

This was achieved by making a scatter plot to visually compare and find correlations between these two variables. The left y-axis corresponds to the range of GDP, which is represented through the yellow line, and the number of suicides shown in purple and this range is indicated by the right y-axis. Through this graph, we found a new story of the causal relationship between numbers of suicide and GDP for year not correlation of those two series. When the GDP graph plunged, the suicide cases increased tremendously, and vice versa. Therefore, we analysed the =RANK function to find out more accurate and deeper analysis.

GDP Table



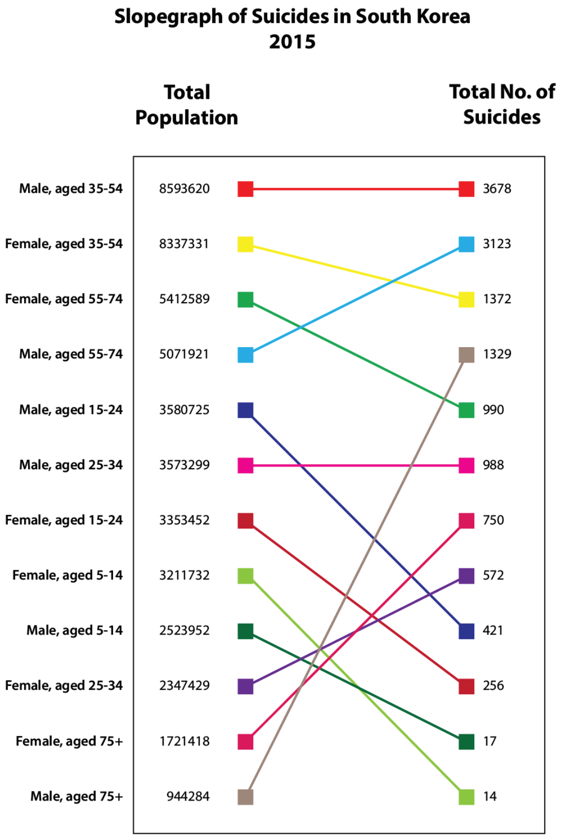
Age	year	suicides_no	Ranking_sucide_no	gdp_for_year ($)	Ranking _GDP
35-54 years	1995	1692	21	556,130,926,913	18
	1996	2065	20	598,099,073,901	15
	1997	2191	19	557,503,074,772	17
	1998	3358	14	374,241,351,752	23
	1999	2770	16	485,248,229,337	20
	2000	2605	18	561,633,125,840	16
	2001	2629	17	533,052,076,314	19
	2002	3289	15	609,020,054,512	14
	2003	4096	12	680,520,724,062	13
	2004	4381	9	764,880,644,711	12
	2005	4288	10	898,137,194,716	11
	2006	3914	13	1,011,797,457,139	8
	2007	4199	11	1,122,679,154,632	6
	2008	4765	8	1,002,219,052,968	9
	2009	5828	2	901,934,953,365	10
	2010	5795	3	1,094,499,338,703	7
	2011	5949	1	1,202,463,682,634	5
	2012	5271	6	1,222,807,284,485	4
	2013	5662	4	1,305,604,981,272	3
	2014	5591	5	1,411,333,926,201	1
	2015	5050	7	1,382,764,027,114	2

The suicide number and GDP for their respective year was ranked, with colour added for easier viewing, with light blue indicating a lower ranking and light orange indicating a higher ranking. In addition to this, the richer orange and blue colours were used to indicate when the rankings of the two variants changed sharply. According to the KDI, 1998, the first year two variables began to make a sharp difference was a period when many people suffered from extreme economic difficulties such as unemployment and bankruptcy (Lee, D., 2016). Comparing the previous year's 1997 suicide with 1998 showed a difference of more than 1,000 people. In addition to this, the early and mid-2000s were also the period of overcoming the Korean financial crisis, but it was the first time that it experienced the credit card crisis, the deepening polarization, and the economic growth rate, which was significantly lower than the previous period. In other words, the economic crisis produced a consistent result of raising the suicide rate of the working population, especially among those aged 35-54.

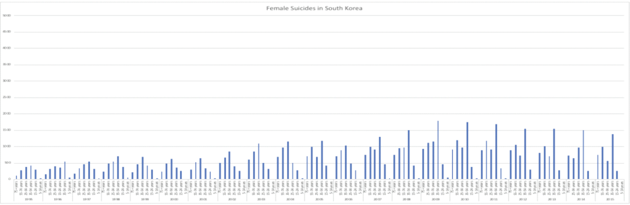
Although this wasn't entirely relevant to our hypothesis, it was still an interesting trend that we explored in our process.

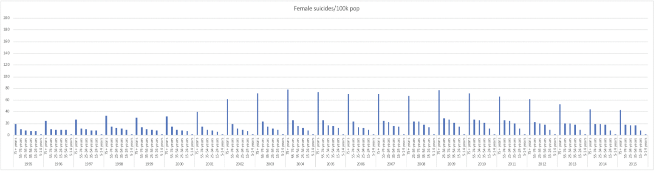
Slope Graph

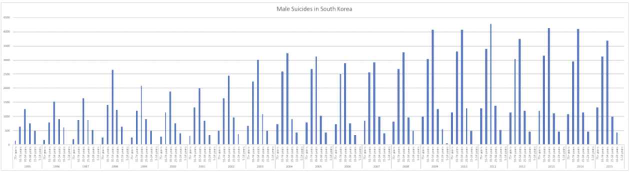
A slopegraph was created as another way to observe the data insightfully. The total population of each year group was compared with their total number of suicides in 2015, revealing surprising information. Those in the aged 75+ category featured relatively high totals of suicide despite having the lowest total population, whilst the youth of the population featured the lowest amount of suicides collectively. The slopegraph was useful for observing the data differently; however the visualization was fairly shallow in terms of the variables it was able to portray, so it did not continue to the final visualization.

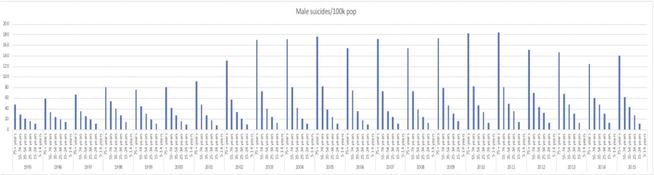


Column Graphs





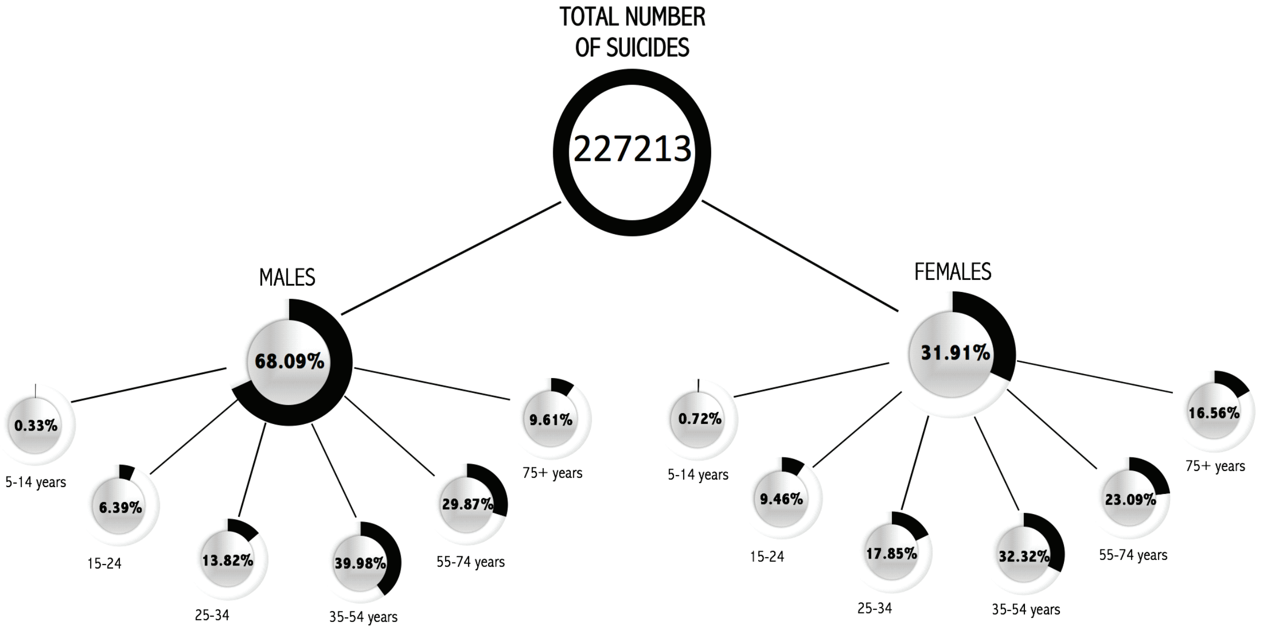




Column graphs were another type of graph explored in the EDA process, and although the absurd amount of information rendered it unsuitable for a potential clear data visualization, it was a helpful reference in visualizing the sheer differences in suicide numbers between age groups and time periods, as well as determining the variables we needed to show in our final visualization. We also determined that the differences in suicide numbers between the two genders required separate graphs as they were quite substantial, with the male graph having a scale twice as large as the female graph.

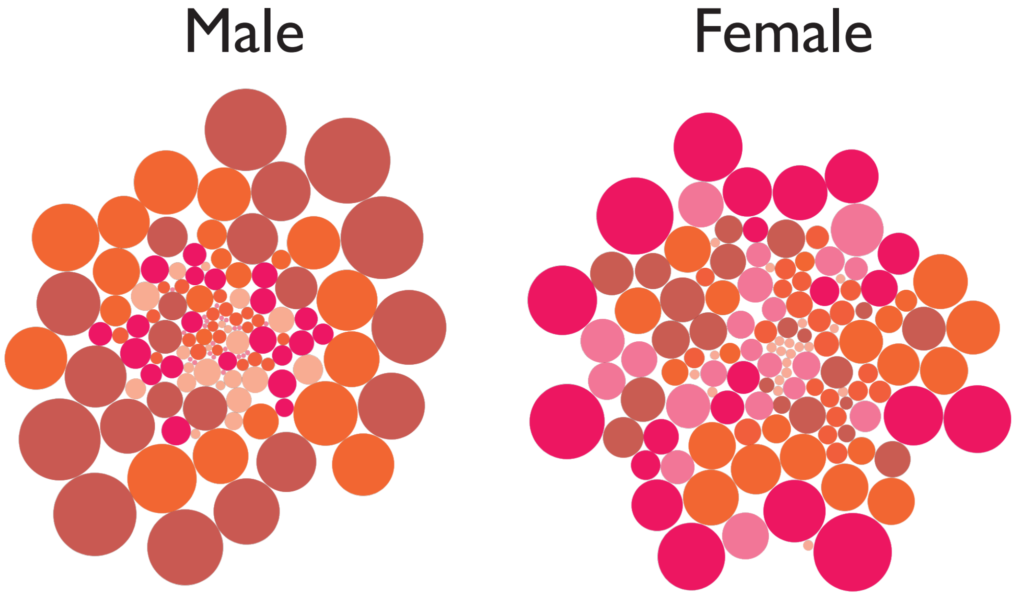
By exploring both of these columns, we've decided to focus on the suicide number per age group rather than the suicides/100k population column for each age group, as representing this on a graph would show an inacurate depiction of the amount that each age group contributes to the total number of deaths

Circle Graph



After this, we decided to use a circle graph in order to investigate more into the age group contributions of each gender. Differentiating from our other graphs, we had also decided to cut the data from before 1995, as there weren't many changes prior and we felt that it could possibly create an unnecessary empty space in our visualisations. We calculated the percentage that each gender contributes to the entire total of deaths, and then from those percentages, we figured out the entire sum of each age group from that gender. We then turned each age group into a percentage of the total of each gender fraction, allowing us to compare each gender more accurately and decipher the trends between both female and male. The 75+ age group accounted to 16.56% of the female deaths, which was almost double as to the male 75+ age group of 9.61%. There was a decrease in the female percentage for both 35-54 and 55-74 years compared to the male percentage. Despite this, the female percentages for ages 5-14, 15-24, 25-35 and 75+ were higher than that of males. We liked this graph as it was more representative of the data and showed the distinctive comparisons between gender and specific age groups, although we also would like to as well demonstrate the growth of each throughout 1995 to 2015. Despite this, we figured that it might be slightly misleading in a sense that the percentages of each age group account for the total overall number of suicides, whereas it's actually a percentage of the total number from each gender.

Gephi Charts

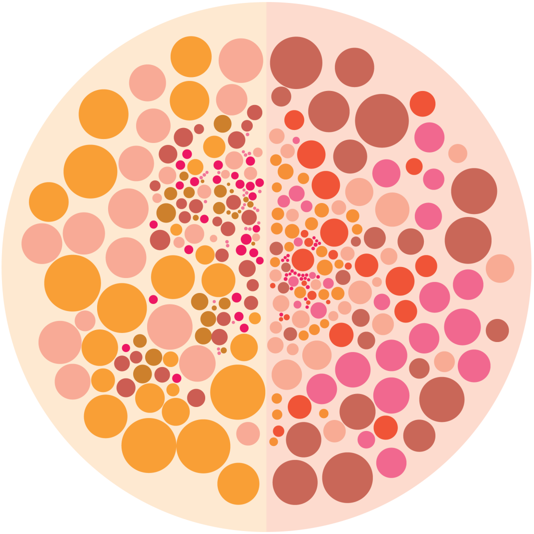


Male

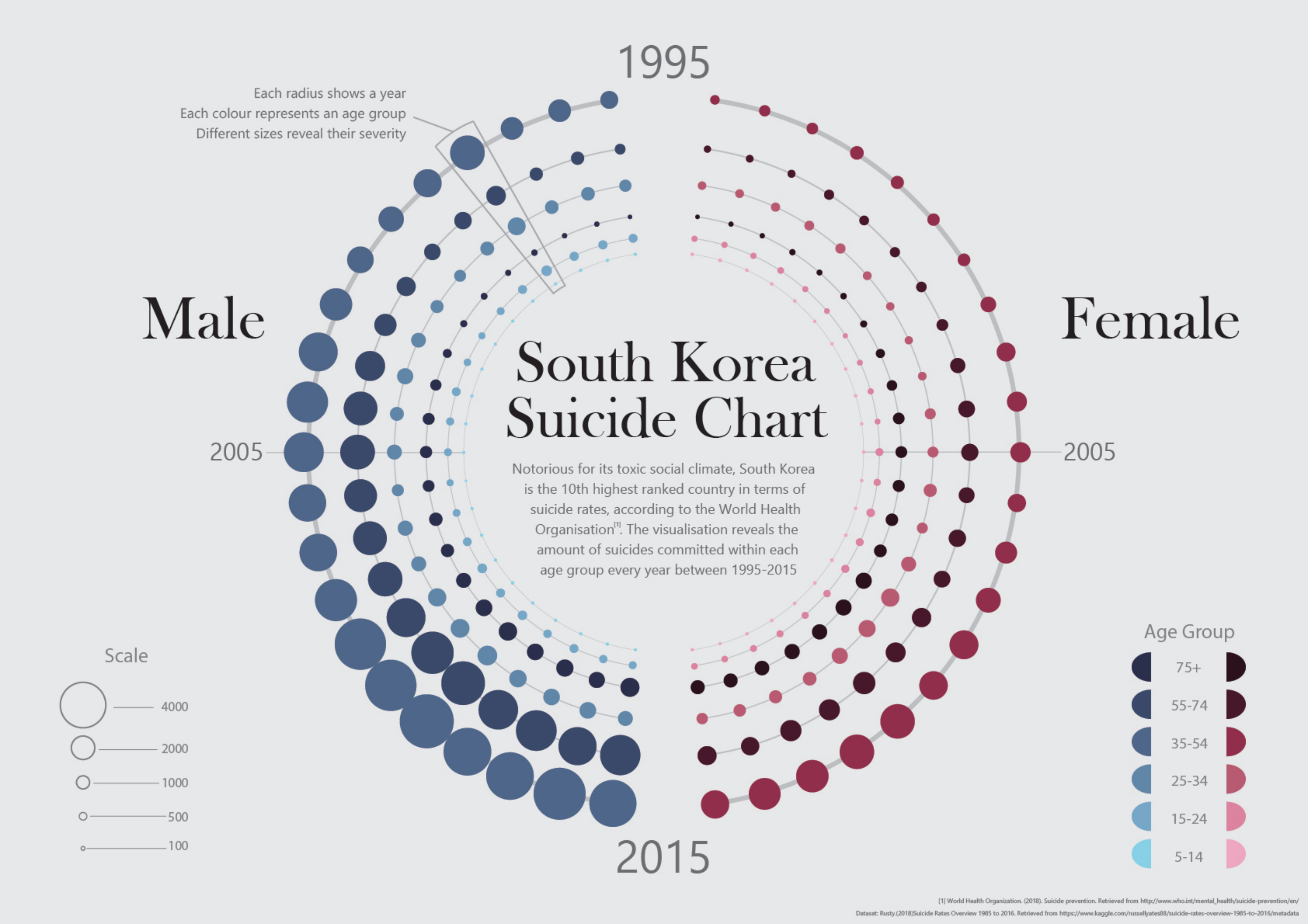
Female

Drawing closer to the final visualization, Gephi charts were made for each gender to explore a more unique method of visualizing the data. In each chart, individual nodes represent particular years, with suicide numbers being attributed to their sizes and age groups differentiated by colours, thus almost all four variables were able to be shown, though the specific year of a particular node would be impossible to distinguish visually. The charts themselves are visually striking and clear, and are effective in their ability to revealing dominant age groups in terms of suicide, as well as conveying the vast difference in numbers with less dominant age groups. As mentioned earlier, year -and as a result, trend- were aspects that were difficult to show in this specific format, and would become a challenge to overcome in the final visualization.

This was an early attempt at creating an appealing visual with the gephi charts above. Although it reflects the same good qualities of the previously shown charts in a more aesthetic form, we decided that year was an important variable to convey, and the lack of chronological organization could be visually overwhelming to audiences.



5. Final Data Visualisation



South Korea Suicide Chart

Notorious for its toxic social climate, South Korea is the 10th highest ranked country in terms of suicide rates, according to the World Health Organisation[1]. The visualisation reveals the amount of suicides committed within each age group every year between 1995-2015

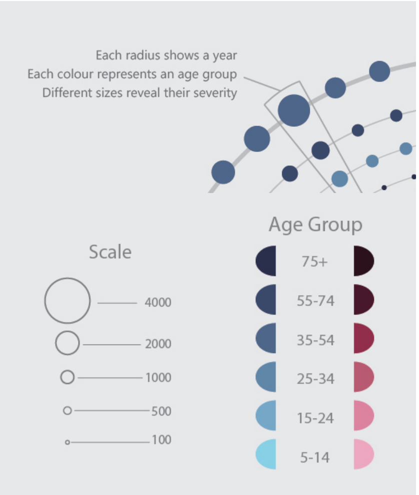
[1] World Health Organization. (2018). Suicide prevention. Retrieved from http://www.who.int/mental_health/suicide-prevention/en/

Dataset: Rusty.(2018)Suicide Rates Overview 1985 to 2016. Retrieved from https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016/metadata

Rationale

Our final visualisation is an iteration of our original gephi chart, revealing data in an approach similar to that of scatter plots. The visualisation conveys four variables with its nodes in distinct ways: gender by set, year by position, age group by colour and suicide numbers by size, altogether revealing the severity of suicide numbers between age groups in South Korea.

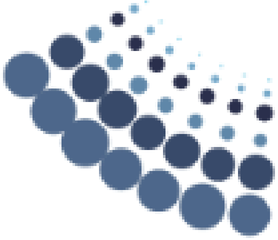
Simple, clear guides



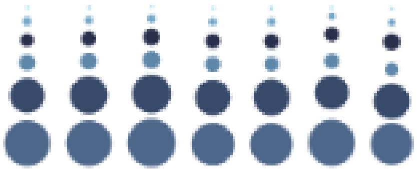
Given the morose subject matter, a minimalistic approach was taken to visualise the data to ensure attention isn't taken away from the chart itself. The narrative and theme of death is subtly contrived by the dull red and blue colours on the light background, which can be likened to bruises and blood clots on a dead body, and perpetuate the melancholy tone of the visualisation (Bosler, n.d.). Textual information is kept relatively scarce, and also given a darker shade of grey to remain unobtrusive from the data. The graph is explained with a single annotation, keeping chart junk to a minimum for a clearer visualisation (Cullen, 2007).

The circular form of the chart serves many functions as a design aspect. It solves the aforementioned challenge of integrating the year variable in an organised fashion, whilst still being a unique, interesting visual, as opposed to a regular scatter plot. It utilises the growing intensity of the nodes to make efficient use of space (refer to the adjacent models), which was a challenge presented by our absurdly large column graph. Whilst in this structure, several principles of Gestalt Theory, such as proximity and similarity, can also be applied to the nodes to distinguish them from each other and allows them to portray each variable clearly (Wertheimer, 1938). As a final point, the concentric nature of the chart acts as a visual vector for the audience, being initially drawn to the title to understand the context, then to the chart around it (Cullen, 2007). The collaboration of these design details results in a visually aesthetic, yet clear and readable final visualisation.

Model 1



Model 2



Conclusion

Through our EDA approach, we discovered that the youth had substantially lower suicide rates compared to those older, thus disproving our hypothesis about the data. However, our exploration led us to explore interesting trends that would correct our preconceived misunderstandings of suicide in Korea, and allowed us to gain a more complete picture. Thus, our final visualisation aims to shed light into the age groups where suicide is most prevalent, and compares them across the years.

6. References

Goh, D. (2019). Gloomy social climate fueling suicide rate in Korea. Retrieved from https://asiatimes.com/2019/06/gloomy-korea-social-climate-fuels-the-rise-of-suicide-rate-in-korea/

Lee, D. (2016). 경기불황이 자살률에 영향을 미칠 수 있을까?. Retrieved from https://eiec.kdi.re.kr/publish/naraView.do?cidx=10777&fbclid=IwAR0MPX0Wlq9vkosqQDeORdyZxe60ygrbxFMlQVI7N1yp25kDjB5QPiXWouY

Lee, S., Park, J., Lee, S., Oh, I., Choi, J., & Oh, C. (2018). Changing trends in suicide rates in South Korea from 1993 to 2016: a descriptive study. BMJ Open, 8(9), e023144. doi: 10.1136/bmjopen-2018-023144

Rusty. (2018). Suicide Rates Overview 1985 to 2016. Retrieved from https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016/metadata

Singh, A. (2017). The “Scourge of South Korea”: Stress and Suicide in Korean Society – Berkeley Political Review. Retrieved from https://bpr.berkeley.edu/2017/10/31/the-scourge-of-south-korea-stress-and-suicide-in-korean-society/

[Szamil]. (2017). Suicide in the Twenty-First Century [dataset]. Retrieved from https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook

United Nations Development Program. (2018). Human development index (HDI). Retrieved from http://hdr.undp.org/en/indicators/137506

World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016. Retrieved from http://databank.worldbank.org/data/source/world-development-indicators#

World Health Organization. (2018). Suicide prevention. Retrieved from http://www.who.int/mental_health/suicide-prevention/en/