FIT5147 DATA EXPLORATION PROJECT

CAUSES OF LIFE SATISFACTION AROUND THE WORLD

STUDENT NAME: KEHAN LIU
STUDENT ID: 32281943
TUTOR: Nina Sophia Dsouza, Tutorial 7

Table of Contents

Introduction *3*

Data wrangling and Data checking.. *3*

Data Checking 3

Data wrangling *4*

Data wrangling for Q1 4

Data wrangling for Q2 4

Data wrangling for Q3 5

Data Exploration *6*

The first 10 countries with highest Life satisfaction 6

How do COVID-19 impact on the life satisfaction in 2020 7

Is working hour correlated with life satisfaction.. 8

Asia vs. Europe 10

Conclusion *12*

Reflection *12*

Reference *13*

Introduction

Life satisfaction now is an important indicator to measure people's life quality, high life satisfaction may have positive effect on people's health even their life span. This report is mainly focus on how the life satisfaction index change in recent year. Then, using both statistical and visualization method to test if there has any relationship between the life satisfaction and coronavirus pandemic as well as working hour. Some conclusion will be given in this report.

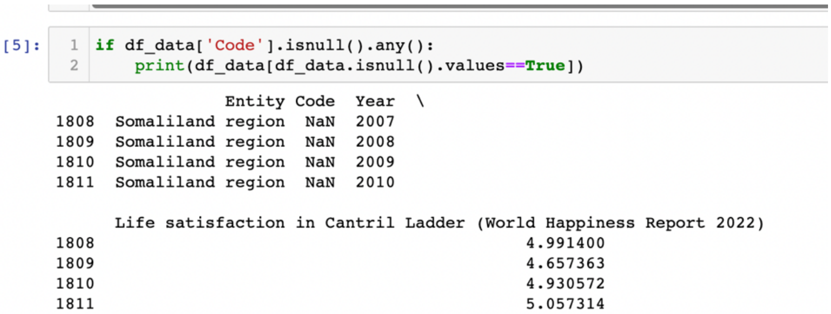
Data wrangling and Data checking

To do the further analysis, the data wrangling and data checking are needed for the raw data of this project. The data wrangling and data checking are done by python with environment python 3.9.7.

df_data: ***happiness-cantril-ladder (1).csv***
df2: ***compare-sources-working-hours.csv***
df_covid: ***COVID-19Cases2020.csv***

Data Checking

This data set used for the visualization project are three csv from different sources, hence there has some problems of the data content (i.e., null values, format of country name). For instance:
Checking the null values, this step is implemented via pandas. The three files have a total of 12,666 null values, some of which are not relevant to this visualization project. The following image shows the method of null values checking. The null value will be removed while using df.dropna() method.



```
[5]:  if df_data['Code'].isnull().any():
          print(df_data[df_data.isnull().values==True])

                  Entity Code  Year  \
1808  Somaliland region  NaN  2007
1809  Somaliland region  NaN  2008
1810  Somaliland region  NaN  2009
1811  Somaliland region  NaN  2010

      Life satisfaction in Cantril Ladder (World Happiness Report 2022)
1808                                           4.991400
1809                                           4.657363
1810                                           4.930572
1811                                           5.057314
```

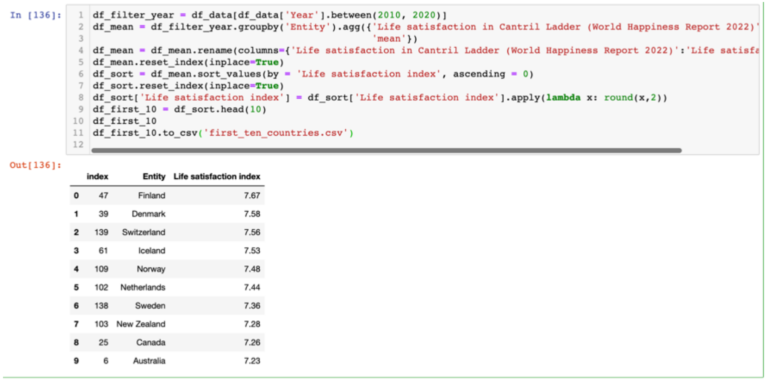
Besides, some countries' names need to change in order to consist with their official name.

For example:
Change the country name “United Republic of Tanzania” to “Tanzania "
Change the country name “Cote d'Ivoire” to “Côte d'Ivoire”
Change the format of country name in “*COVID-19Cases2020.csv*”. Some countries names with more than one word in this file are connected with underscore, replace the underscore with whit space.

Data wrangling

Data wrangling for Q1

Filter data between 2010 and 2020, and group by Entity name and aggregate with mean value of life satisfaction index. Sorting the data and get the first 10 countries data. The out put file is *“first_ten_countries.csv”*.



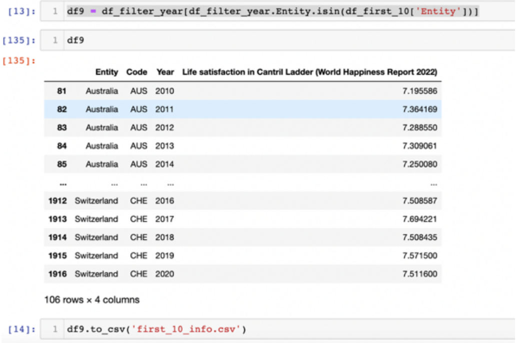
In [136]:

```
df_filter_year = df_data[df_data['Year'].between(2010, 2020)]
df_mean = df_filter_year.groupby('Entity').agg({'Life satisfaction in Cantril Ladder (World Happiness Report 2022)'
                                                 'mean'})
df_mean = df_mean.rename(columns={'Life satisfaction in Cantril Ladder (World Happiness Report 2022)':'Life satisfa
df_mean.reset_index(inplace=True)
df_sort = df_mean.sort_values(by = 'Life satisfaction index', ascending = 0)
df_sort.reset_index(inplace=True)
df_sort['Life satisfaction index'] = df_sort['Life satisfaction index'].apply(lambda x: round(x,2))
df_first_10 = df_sort.head(10)
df_first_10
df_first_10.to_csv('first_ten_countries.csv')

```

Out[136]:

	index	Entity	Life satisfaction index
0	47	Finland	7.67
1	39	Denmark	7.58
2	139	Switzerland	7.56
3	61	Iceland	7.53
4	109	Norway	7.48
5	102	Netherlands	7.44
6	138	Sweden	7.36
7	103	New Zealand	7.28
8	25	Canada	7.26
9	6	Australia	7.23

The details information of the first 10 countries are got by the following code. Output file with “first_10_info.csv”



[13]:

```
df9 = df_filter_year[df_filter_year.Entity.isin(df_first_10['Entity'])]
```

[135]:

```
df9
```

[135]:

	Entity	Code	Year	Life satisfaction in Cantril Ladder (World Happiness Report 2022)
81	Australia	AUS	2010	7.195586
82	Australia	AUS	2011	7.364169
83	Australia	AUS	2012	7.288550
84	Australia	AUS	2013	7.309061
85	Australia	AUS	2014	7.250080
...	...	...	...	...
1912	Switzerland	CHE	2016	7.508587
1913	Switzerland	CHE	2017	7.694221
1914	Switzerland	CHE	2018	7.508435
1915	Switzerland	CHE	2019	7.571500
1916	Switzerland	CHE	2020	7.511600

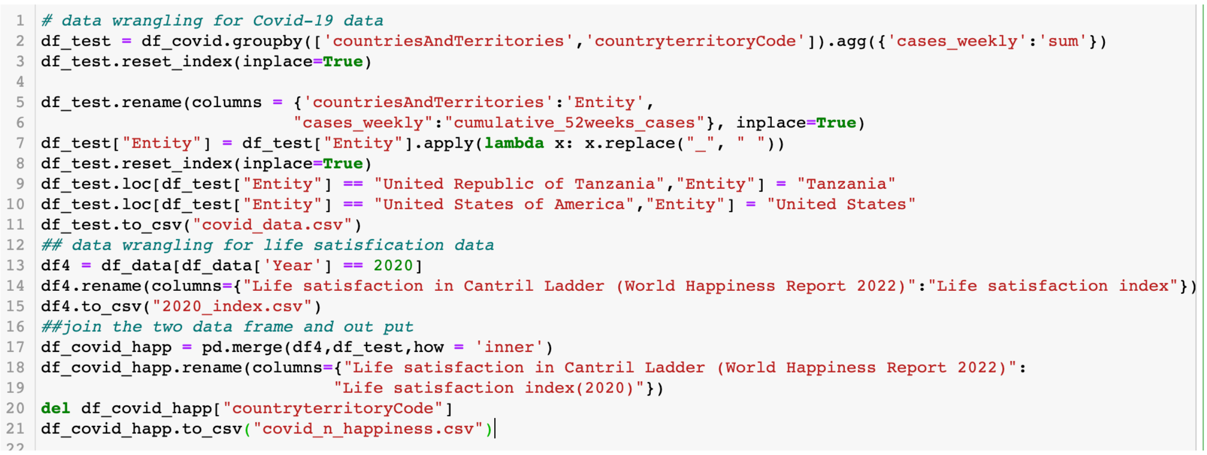
106 rows × 4 columns

[14]:

```
df9.to_csv('first_10_info.csv')
```

Data wrangling for Q2

Reformat the country name in raw COVID-19 data set
Merge the two csv file, output file name as *”covid_n_happiness.csv”*



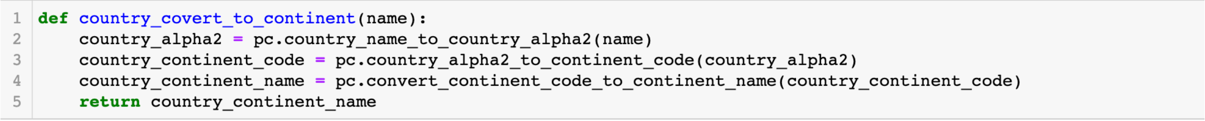
```python
# data wrangling for Covid-19 data
df_test = df_covid.groupby(['countriesAndTerritories','countryterritoryCode']).agg({'cases_weekly':'sum'})
df_test.reset_index(inplace=True)

df_test.rename(columns = {'countriesAndTerritories':'Entity',
                         "cases_weekly":"cumulative_52weeks_cases"}, inplace=True)
df_test["Entity"] = df_test["Entity"].apply(lambda x: x.replace("_", " "))
df_test.reset_index(inplace=True)
df_test.loc[df_test["Entity"] == "United Republic of Tanzania","Entity"] = "Tanzania"
df_test.loc[df_test["Entity"] == "United States of America","Entity"] = "United States"
df_test.to_csv("covid_data.csv")
## data wrangling for life satisfication data
df4 = df_data[df_data['Year'] == 2020]
df4.rename(columns={"Life satisfaction in Cantril Ladder (World Happiness Report 2022)":"Life satisfaction index"})
df4.to_csv("2020_index.csv")
##join the two data frame and out put
df_covid_happ = pd.merge(df4,df_test,how = 'inner')
df_covid_happ.rename(columns={"Life satisfaction in Cantril Ladder (World Happiness Report 2022)":
                             "Life satisfaction index(2020)"})
del df_covid_happ["countryterritoryCode"]
df_covid_happ.to_csv("covid_n_happiness.csv")
```

Data wrangling for Q3

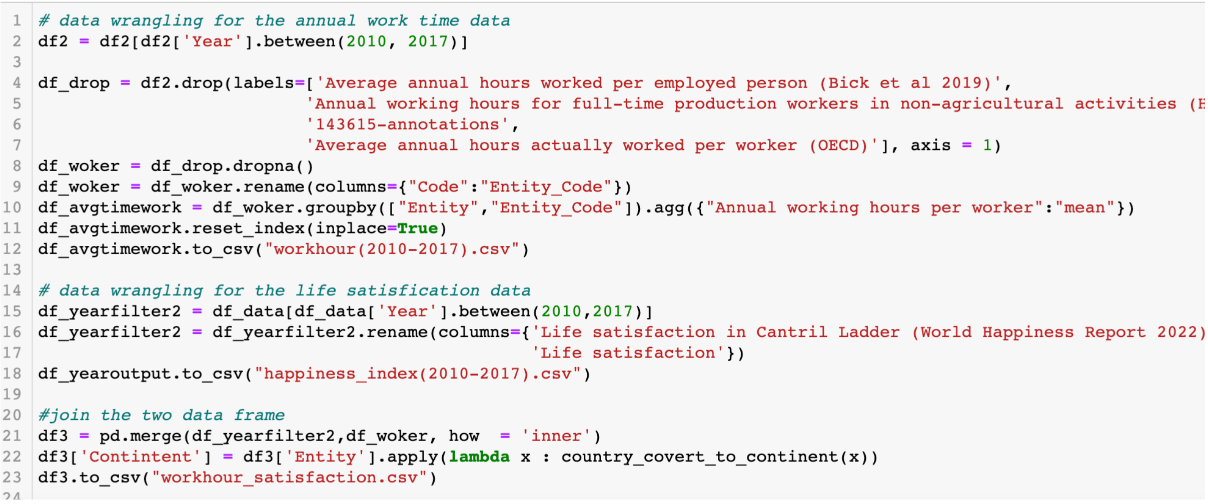
Define the function to get the Continent name of each country

Create a new column for continent

Merge the two files, output file name *"workhour_satisfaction.csv"*



```python
def country_covert_to_continent(name):
    country_alpha2 = pc.country_name_to_country_alpha2(name)
    country_continent_code = pc.country_alpha2_to_continent_code(country_alpha2)
    country_continent_name = pc.convert_continent_code_to_continent_name(country_continent_code)
    return country_continent_name
```



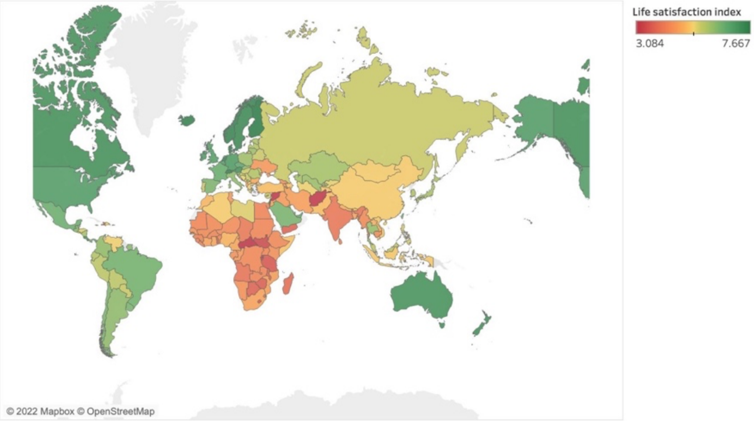
```python
# data wrangling for the annual work time data
df2 = df2[df2['Year'].between(2010, 2017)]

df_drop = df2.drop(labels=['Average annual hours worked per employed person (Bick et al 2019)',
                           'Annual working hours for full-time production workers in non-agricultural activities (E
                           '143615-annotations',
                           'Average annual hours actually worked per worker (OECD)'], axis = 1)
df_woker = df_drop.dropna()
df_woker = df_woker.rename(columns={"Code":"Entity_Code"})
df_avgtimework = df_woker.groupby(["Entity","Entity_Code"]).agg({"Annual working hours per worker":"mean"})
df_avgtimework.reset_index(inplace=True)
df_avgtimework.to_csv("workhour(2010-2017).csv")

# data wrangling for the life satisfication data
df_yearfilter2 = df_data[df_data['Year'].between(2010,2017)]
df_yearfilter2 = df_yearfilter2.rename(columns={'Life satisfaction in Cantril Ladder (World Happiness Report 2022)
                                                'Life satisfaction'})
df_yearoutput.to_csv("happiness_index(2010-2017).csv")

#join the two data frame
df3 = pd.merge(df_yearfilter2,df_woker, how  = 'inner')
df3['Contintent'] = df3['Entity'].apply(lambda x : country_covert_to_continent(x))
df3.to_csv("workhour_satisfaction.csv")
```

Data Exploration

The first 10 countries with highest Life satisfaction

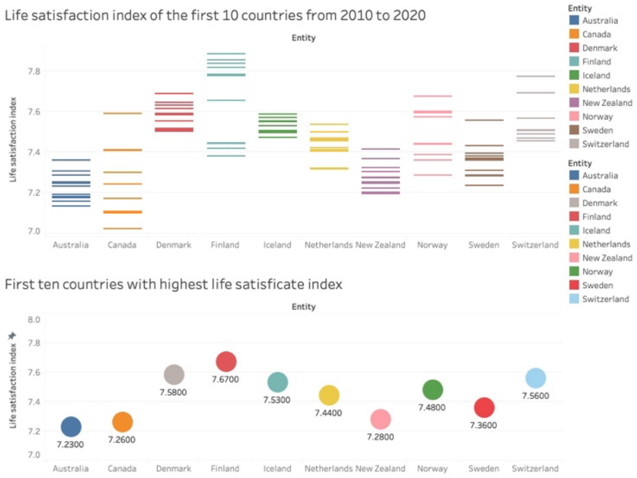
Life satisfaction now become an important topic with people's daily life. Life satisfaction refers to how much a person enjoys his or her life(Tokay& Mersin, 2021). The higher life satisfaction may have a positive effect on individual's life quality even their expectation of life span.

The figure below shows the overview of the life satisfaction index around the world. The life index is calculated by the average value of life index from 2010 to 2020.



Map based on Longitude (generated) and Latitude (generated). Color shows the avergae number of Life satisfaction index from 2010 to 2020. Details are shown for Entity. The view is filtered on Latitude (generated) and Longitude (generated). The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only.

The charts below shows the first ten countries with highest life satisfaction index. Overall, these countries all have a happiness index of around 7.5. And the distribution of these countries' life satisfaction indices between 2010 and 2020 is regular, concentrated in a certain range or increasing. Besides, if we observe these countries' names we can find that these countries are all belongs to the developed countries and this could explain why these countries have relatively high life satisfaction indices; because they have comprehensive public facilities and better social welfare, both of which have a direct impact on individual's satisfaction with their life.



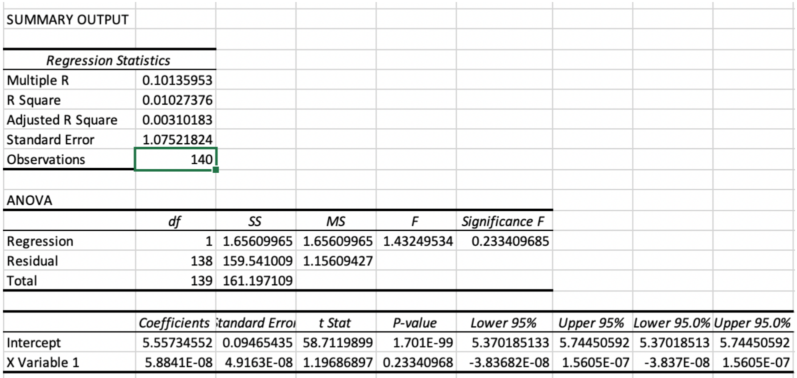
An overview of the first ten countries'life index information. color shows entities. The chart at the top of dashborad shows the life satisfaction index from 2010 to 2020 of the first 10 countries with highest average life index in past 10 years. And the grah at the bottom of the dashboard shows shows the average life index between 2010 and 2020

How do COVID-19 impact on the life satisfaction in 2020

COVID-19 broke out in early 2020 and has had a significant impact on the lives of many people. To explore whether the COVID-19 is impact on people's life satisfaction or not, we will run the hypothesise test (with $\alpha = 5\%$) in this part.

H_0: COVID-19 does not effect on people's life satisfaction in 2020
H_1: COVID-19 has effect on people's life satisfaction in 2020

The regression output shows as below. The X variable is the cumulative COVID-19 cases within 52 weeks in 2020 of each countries. Based on the result, the p-value is around 0.233 which caused fail to reject the H_0 under the 95% confidence interval. And this result shows that there is no linear relationship between the COVID-19 cases and life satisfaction index, in other word, the COVID-19 may not have very significant impact on people's life satisfaction.



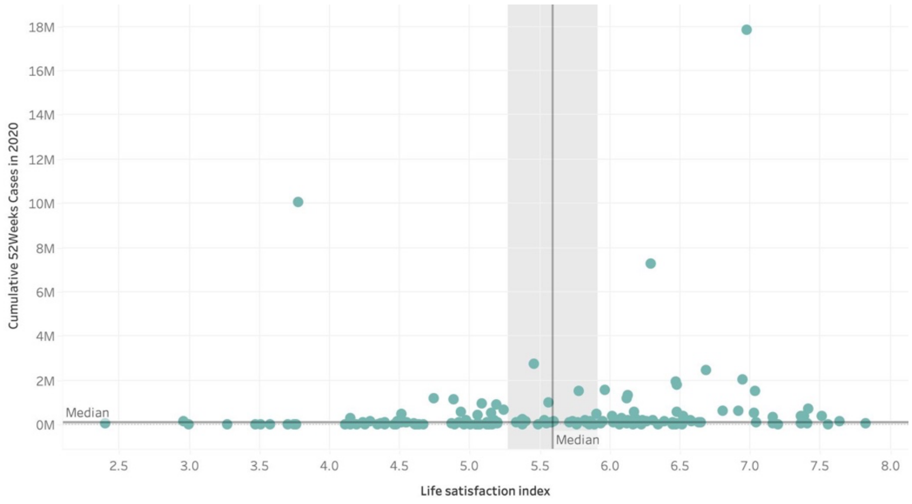
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.10135953
R Square	0.01027376
Adjusted R Square	0.00310183
Standard Error	1.07521824
Observations	140

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.65609965	1.65609965	1.43249534	0.233409685
Residual	138	159.541009	1.15609427		
Total	139	161.197109			

	Coefficients	*tandard Erroi*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	5.55734552	0.09465435	58.7119899	1.701E-99	5.370185133	5.74450592	5.37018513	5.74450592
X Variable 1	5.8841E-08	4.9163E-08	1.19686897	0.23340968	-3.83682E-08	1.5605E-07	-3.837E-08	1.5605E-07



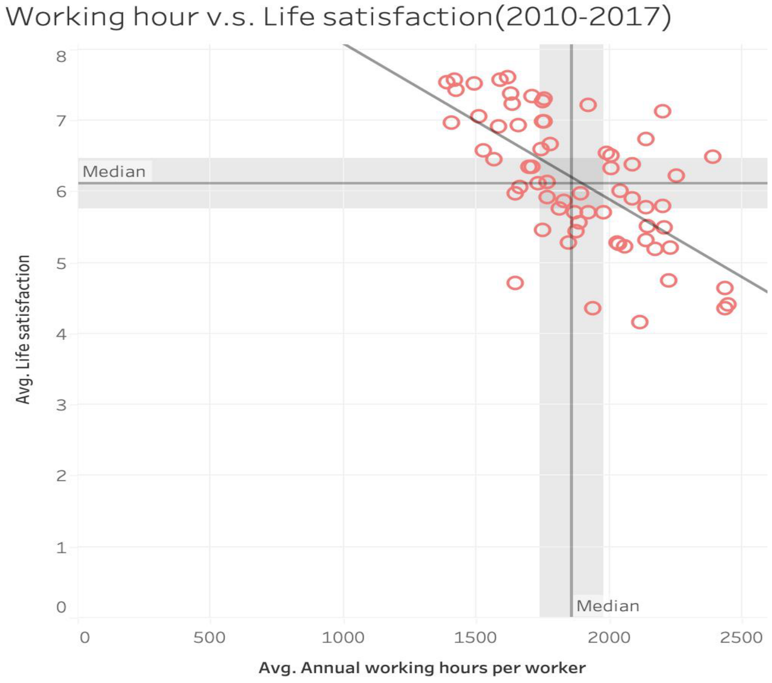
Life satisfaction index vs. Cumulative 52Weeks Cases. Details are shown for Entity.

By visualizing the data, we can find that more than half of observations' has relatively low life satisfaction index even though they has low cumulative confirmed COVID-19 cases (i.e., lower than one million) in 2020. On contrary, some countries have high cumulative confirmed cases but their life satisfaction index still at a higher figure.

Is working hour correlated with life satisfaction

Work-life balance has always been the goal that people have sought. Is there really a correlation between working hours ad individual's life satisfaction? Does the more work people do, the happier they are? The visualization below illustrate the relationship between the life satisfaction and annual working hour. The trend line indicate there has a negative relationship between life satisfaction index and working hour. The line regression model between these variable can be written as:

$$Avg. Life\ satisfaction = -0.00219264 * Avg\ Annual\ working\ hour + 10.2714$$

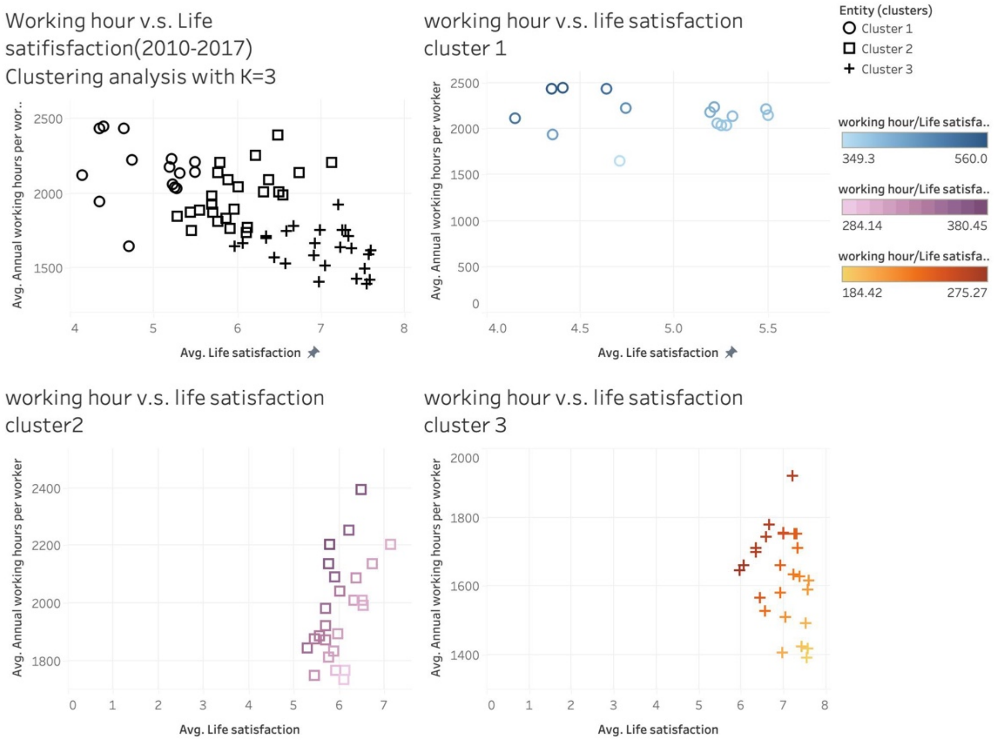


Average of Annual working hours per worker vs. average of Life satisfaction. Details are shown for Entity.

The charts below show the cluster analysis of working hour and life satisfaction, with K = 3. We define a working-life ratio in this part's analysis:

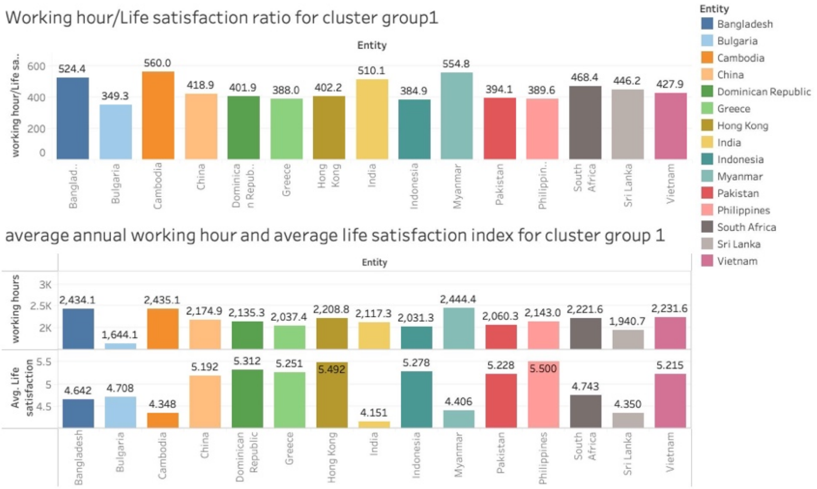
Working-life ratio = Average annual working hour / Life satisfaction index

Assuming the working hour is relatively fixed, the higher ratio means that this country's this country's working-life balance cannot be achieved well since they have low life satisfaction index.



An overview of clutering analysis. shape shows the different group of cluster. color shows the working hour/Life satisfaction

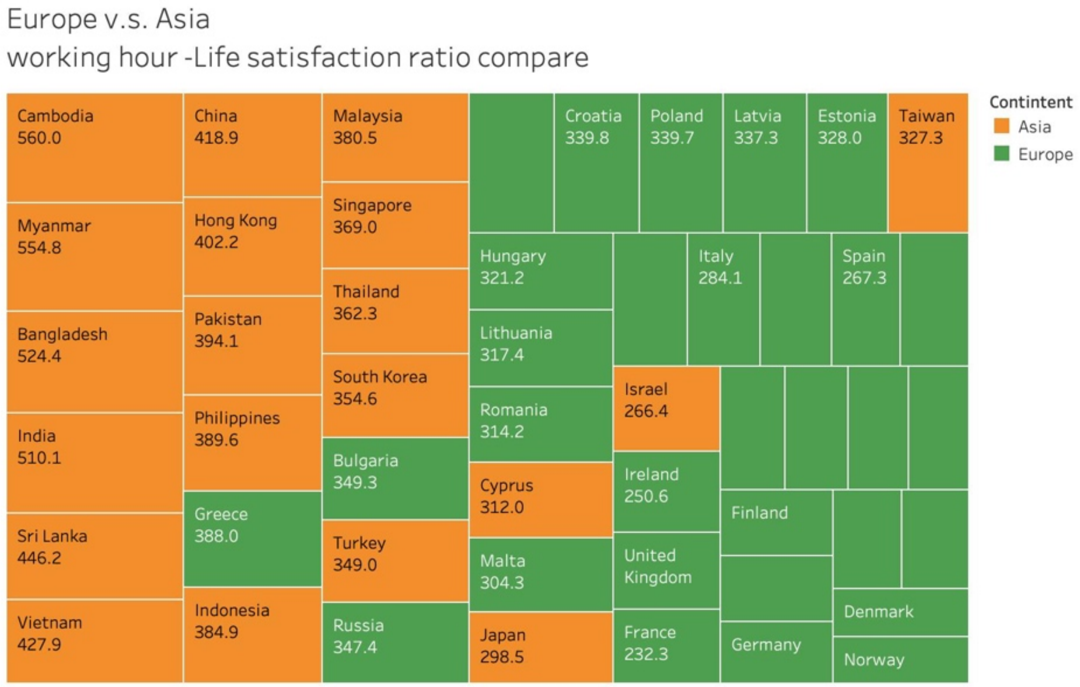
Taking the cluster1 as example, this group has relatively large scale of working hour-life satisfaction ratio among these cluster group.



By observing the detail information in this cluster group above, we can find that most of countries in this group are developing countries, and this might be a reason why they have high figure of annual working hour per worker but low life satisfaction index. Workers in these countries need to work more to contribute to their economies, but on the other hand, social welfare in these countries is not as generous as in developed countries. Hence, people in these countries live under more stress than the developed countries.

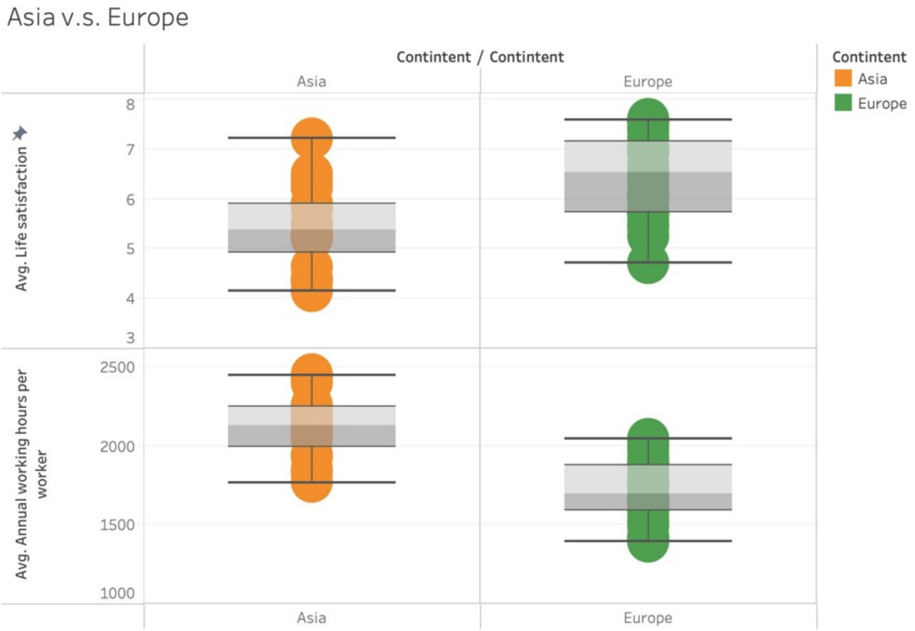
Asia vs. Europe

This section we will compare the analysis result within Asia and Europe. The figure below shows the working hour-Life satisfaction ratio in Asia and European countries. Based on the graph, we can see that most countries or region in Asia has high work-life ratio, by contrast, the corresponding figure in European countries are relatively low, around 250.



Entity and working hour/Life satisfaction. Color shows details about Contintent. Size shows working hour/Life satisfaction ratio. The marks are labeled by Entity and working hour/Life satisfaction. The view is filtered on Contintent, which keeps Asia and Europe.

The graph above shows indicate that compare with Asian countries, most entities in Europe has low annual working hour per worker but high life satisfaction index. Is that true? The following visualization illustrate average annual working hours and average life satisfaction index in Asian and European countries. The figure indicate that the median of average annual working hour in European countries is lower than the Asian countries counterpart (1670hours vs. 2130hours respectively), but has higher average life satisfaction index (6.54 vs. 5.38 respectively).



Average of Life satisfaction and average of Annual working hours per worker for each Contintent broken down by Contintent. Color shows details about Contintent. Details are shown for Contintent and Entity. The view is filtered on Contintent, which keeps Asia and Europe.

Conclusion

In summary, the life satisfaction is affected by many factors in society or people's daily life. This project select two popular factors in recent year and the result are completely different. By doing the hypothesis testing of the relationship between COVID-19 cases in 2020 and life satisfaction index in 2020, we found that the COVID-19 pandemic does not have significant effect on people's life satisfaction in 2020. On the contrary, we found that there has a negative relationship in working hour and life satisfaction.

Reflection

People's satisfaction with their lives is a subjective attitude, which is influenced by many real and objective factors. This project only select two hot topic in recent year to explore if there has some relationship between people's life satisfaction and COVID-19 as well as working hour. The result may inaccuracy. Besides, there has missing data of some countries (i.e. North Korea, etc.). And these missing value may cause the error of evaluation.

Reference

Tokay Argan, Mehpare, & Mersin, Sevinç (2021). Life satisfaction, life quality, and leisure satisfaction in health professionals. *Perspectives in Psychiatric Care*, 57(2), 660–666. https://doi.org/10.1111/ppc.12592

Data source:

Life satisfaction in cantril ladder, world happiness report 2022. Retrieved from: https://ourworldindata.org/grapher/happiness-cantril-ladder

COVID-19 historical data to 14 December, 2020. Retrieved from: https://ourworldindata.org/grapher/annual-working-hours-per-worker?tab=table

Annual working hours per worker, published by Huberman & Minns (2007); PWT 9.1 (2019). Retrieved from: https://www.kaggle.com/datasets/fedesoriano/covid19-historical-data-to-14-december-2020