Econ 406 final project write up

The dataset I used in the final project is about world happiness report from Kaggle (https://www.kaggle.com/mathurinache/world-happiness-report).

Introduction: The World Happiness report is a worldwide survey contains the happiness score for 153 countries with factors used to explain the score. It proceeds to gain worldwide acknowledgement as government, organizations, and society progressively use happiness index to refine their policy and making critical decisions.

Content: From the website: "World Happiness Report is ranked among 156 Countries based on a Cantril Ladder Survey. Nationally representative samples of respondents are asked to think of a ladder, the best possible life for them being a 10, and the worst possible experience is a 0. They are then asked to rate their own current lives on that 0 to 10 scale. The columns taking after the bliss score assess the degree to which each of six variables – financial generation, social back, life anticipation, flexibility, nonattendance of debasement, and liberality – contribute to making life assessments higher in each nation than they are in Dystopia, a theoretical nation that has values rise to the world's least national midpoints for each of the six variables."

Analysis of World Happiness Report

I was interested in the following question: which of the factor/factors in the dataset most correlated to the happiness level?

I think this question is important because the government, the public, society can focus on making policies and educating themselves to improve those factors, thus ameliorate the happiness level of their countries. Also, we can reflect on how to help those countries with low happiness level to make the world a better place.

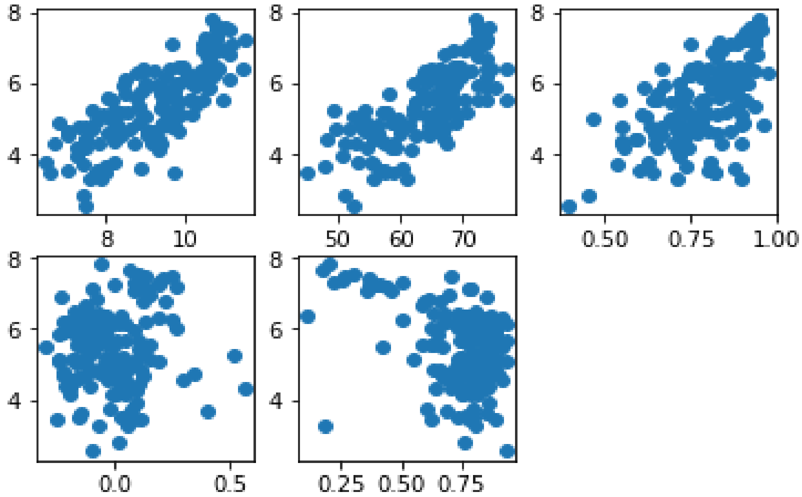
I will investigate some factors correlated to the happiness level including life-expectancy, GDP per capita, gender equality, etc. I will draw plot and modeling those independent variables to investigate their correlations to happiness level.

1. First step of doing data analysis is to clean the dataset. After importing the dataset, clean the data by drop NA values. This ensures that I have the dataset with useful values for later prediction and modeling. I did this in the first function named **import_and_clean_data**.

2. Next, I want to discover the strength of correlation between different factors to the happiness level. I utilize GDP per Capita, Health Life Expectancy, Freedom to make choices, Government Corruption, and Generosity as five independent variables

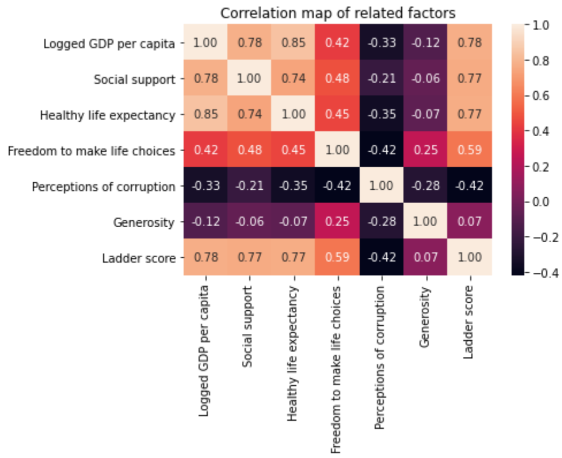
regarding Happiness Score.

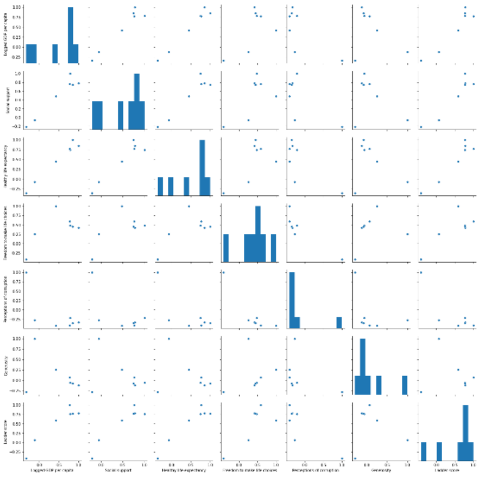
I used three different plots in this section. The first one is the scatter plot, it allows me to infer the strength of correlation directly by looking the dots on the plot.



Conclusion: The first two is GDP per capita and Healthy life expectancy corresponding to happiness level, which present us strong correlation as the dots concentrate along a invisible line, so we can tell that these two factors mostly correlated to the happiness level. The second is Freedom to make life choices, which also correlate but not strong as the former two. The third one is Generosity, we can see the dots scatter through the plot which means it has much weak correlation then the previous plots. The fifth is Perceptions of corruption, it is scattered but we can tell a negative correlation, which means the more corrupt the government is, the lower the happiness level for the country.

The second is the heatmap. As I'm interested in the correlation between each metric in the dataset. The third one is the pairplot, it also shows the correlation between each metric but in different way.





Conclusion: from these two plot we can not only get the correlation between each different factor, but can further confirm that the Happiness Score is correlated with

GDP per capita, Health life expectancy, and freedom, but is not correlated to Generosity.

3. I wrote a function to generate descriptive statistics for the dataset.

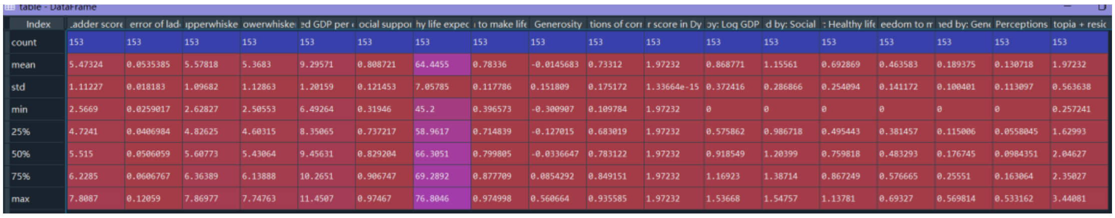
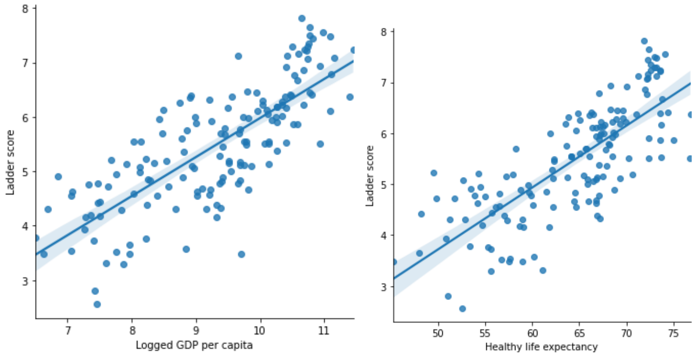


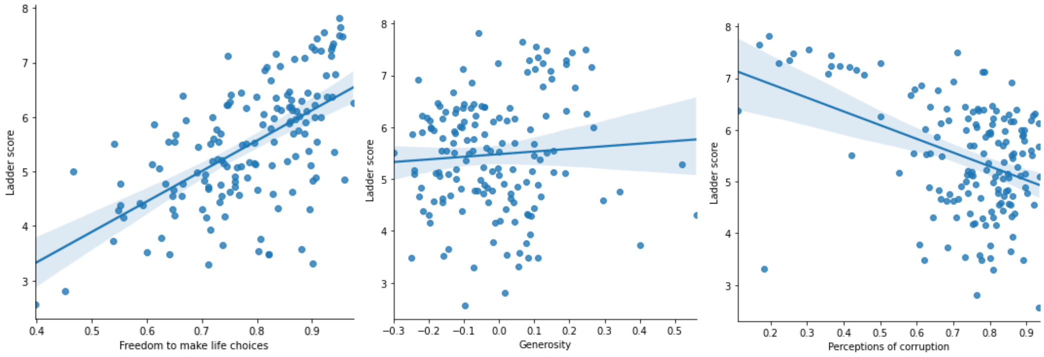
table - DataFrame

Index	.adder score	error of lad	ıpperwhiske	owerwhiske	ed GDP per	ocial suppor	hy life expec	to make life	Generosity	tions of corr	r score in Dy	by: Log GDP	d by: Social	: Healthy life	eedom to m	ıed by: Gene	Perceptions	topia + resic
count	153	153	153	153	153	153	153	153	153	153	153	153	153	153	153	153	153	153
mean	5.47324	0.0535385	5.57818	5.3683	9.29571	0.808721	64.4455	0.78336	-0.0145683	0.73312	1.97232	0.868771	1.15561	0.692869	0.463583	0.189375	0.130718	1.97232
std	1.11227	0.018183	1.09682	1.12863	1.20159	0.121453	7.05785	0.117786	0.151809	0.175172	1.33664e-15	0.372416	0.286866	0.254094	0.141172	0.100401	0.113097	0.563638
min	2.5669	0.0259017	2.62827	2.50553	6.49264	0.31946	45.2	0.396573	-0.300907	0.109784	1.97232	0	0	0	0	0	0	0.257241
25%	4.7241	0.0406984	4.82625	4.60315	8.35065	0.737217	58.9617	0.714839	-0.127015	0.683019	1.97232	0.575862	0.986718	0.495443	0.381457	0.115006	0.0558045	1.62993
50%	5.515	0.0506059	5.60773	5.43064	9.45631	0.829204	66.3051	0.799805	-0.0336647	0.783122	1.97232	0.918549	1.20399	0.759818	0.483293	0.176745	0.0984351	2.04627
75%	6.2285	0.0606767	6.36389	6.13888	10.2651	0.906747	69.2892	0.877709	0.0854292	0.849151	1.97232	1.16923	1.38714	0.867249	0.576665	0.25551	0.163064	2.35027
max	7.8087	0.12059	7.86977	7.74763	11.4507	0.97467	76.8046	0.974998	0.560664	0.935585	1.97232	1.53668	1.54757	1.13781	0.69327	0.569814	0.533162	3.44081

Conclusion: We can tell from this summary table that health life expectancy has the largest standard deviation means countries from different region have hugely different life-expectancy. The government should focus on improving life-expectancy by avoiding war, hunger, disease to help with the happiness level. Critically speaking, we should reevaluate what are the things that we should prioritize in our lives, and how material and spiritual support can improve our well-beings.

4. I plotted the regression line and chose to use the OLS model since the intercept didn't go below zero. I wrote the formula for the model with happiness on the left side and the other variables on the right side, and generate the OLS table to convince my observation from the scatterplot.

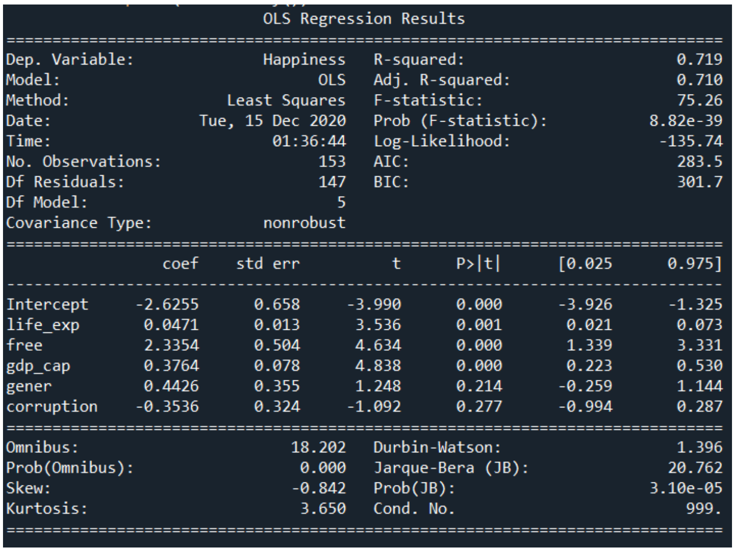




Conclusion: We can tell from the plot with regression line that GDP per capita, health life expectancy, and freedom to make life choices are strongly correlated with the Ladder score (happiness score). Generosity displays almost no correlation with the happiness level, and the perception of government corruption displays negative correlation with the happiness level.

The model I used is Happiness~life_exp+free+gdp_cap+gener+corruption, taking five factors into consideration, and investigate which of them are correlated to the happiness level and which are not.

The function also generates the OLS regression table.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Happiness   R-squared:                       0.719
Model:                            OLS   Adj. R-squared:                  0.710
Method:                 Least Squares   F-statistic:                     75.26
Date:                Tue, 15 Dec 2020   Prob (F-statistic):           8.82e-39
Time:                        01:36:44   Log-Likelihood:                -135.74
No. Observations:                 153   AIC:                             283.5
Df Residuals:                     147   BIC:                             301.7
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -2.6255      0.658     -3.990      0.000      -3.926      -1.325
life_exp       0.0471      0.013      3.536      0.001       0.021       0.073
free           2.3354      0.504      4.634      0.000       1.339       3.331
gdp_cap        0.3764      0.078      4.838      0.000       0.223       0.530
gener          0.4426      0.355      1.248      0.214      -0.259       1.144
corruption    -0.3536      0.324     -1.092      0.277      -0.994       0.287
==============================================================================
Omnibus:                       18.202   Durbin-Watson:                   1.396
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               20.762
Skew:                          -0.842   Prob(JB):                     3.10e-05
Kurtosis:                       3.650   Cond. No.                         999.
==============================================================================
```

From the table, we know that the p value for life expectancy, freedom of choices, GDP per capita are all smaller than 0.05, which means it’s less than 0.05 percent hance we are wrong to reject there is no correlation of life expectancy, freedom of choices, GDP per capita to happiness level. This means that all these three factors are highly correlated to the happiness level.

The coefficient of life expectancy, freedom of choices, and GDP per capita are all positive, which means they are all positively correlated to the happiness level, while the government corruption is negatively correlated to the happiness level.

The R-squared helps to explain how well the model of prediction. The R-squared is 0.719 here which means there is 71.9% of the data fits the regression model, and indicating more than half of the data of happiness can be explained by the model created regarding those five variables.

Overall conclusion:

1. Happiest countries have higher GDP per capita, life expectancy and people have more freedom to make choices compared to unhappy countries.

2. There is a vast difference (big standard deviation) in the life expectancy of different countries, which also being a big factor impacting the happiness of the countries.