Project-1

Saniya Abushakimova

2/26/2021

Introduction

The purpose of this project is to explore the relationship between Life Satisfaction and GDP per capita. The project is particularly interested in answering the following question: *"Did the life satisfaction of people living in European countries in 2017 depend on the country's GDP?"*

For the purpose of this analysis, I used the dataset from Our World in Data (https://ourworldindata.org/grapher/gdp-vs-happiness). The dataset describes 287 countries by giving information about GDP per capita and the life satisfaction of these countries in the time span from 2005 to 2017.
X: GDP per capita is gross domestic product converted to *international dollars*. It was collected from World Bank, International Comparison Program database (http://data.worldbank.org/data-catalog/world-development-indicators).
Y: The life satisfaction rate was collected by Gallup World Poll surveys published in World Happiness Report 2019 (https://worldhappiness.report/ed/2019/). It illustrates the average of survey responses to the 'Cantril Ladder' question, in which the best possible life is rated as 10, and the worst one as 0.

This project will focus only on European countries in 2017. In order to narrow down the dataset, I performed some preprocessing using *Python (pandas)*. Preprocessing procedures included dropping null values and unnecessary columns (Code, Total population, Continent), renaming some columns, and sorting out the dataset to European countries in 2017. More details can be found in *Preprocessing/preprocessing.ipynb*.

The first six records of the dataset are shown below (notice that GDP is given in $):

```
# Loading and printing the data
library(knitr)
data <- data.frame(read.csv(file ='data.csv'))
attach(data)
kable(head(data),align = 'c')
```

Country	Year	Life_satisfaction	GDP
Albania	2017	4.639548	11803.43
Austria	2017	7.293728	45436.69
Belarus	2017	5.552915	17167.97
Belgium	2017	6.928348	42658.58
Bulgaria	2017	5.096902	18563.31
Croatia	2017	5.343166	22669.80

Exploratory data analysis

Let's start the exploratory data analysis from a **statistical summary** in order to get a general understanding about the data.

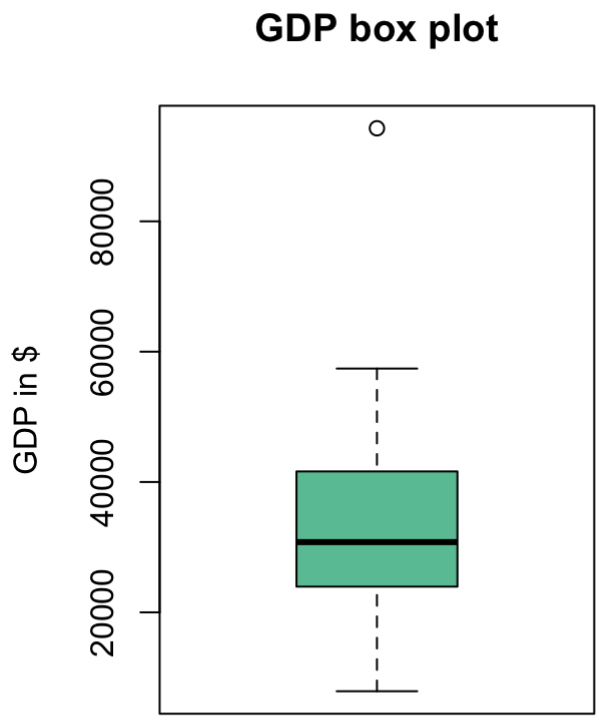
```
summary(data)
```

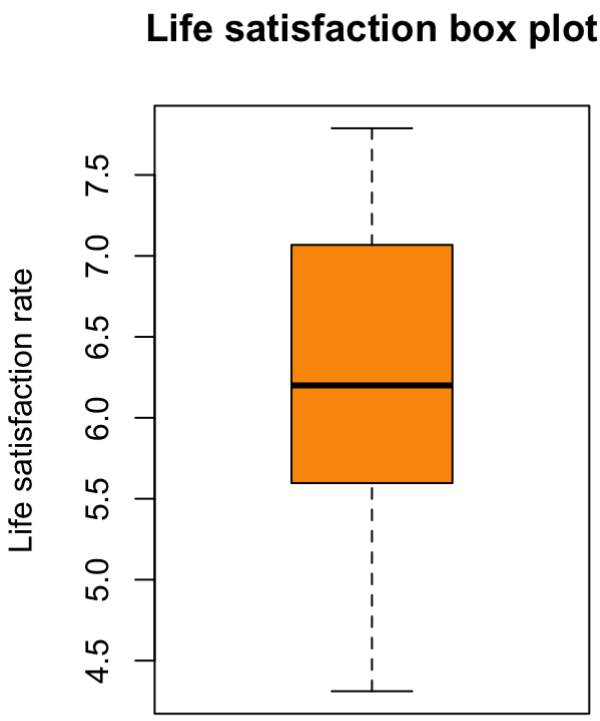
```
##    Country               Year      Life_satisfaction      GDP
##  Length:36          Min.   :2017   Min.   :4.311     Min.   : 7894
##  Class :character   1st Qu.:2017   1st Qu.:5.606     1st Qu.:24259
##  Mode  :character   Median :2017   Median :6.200     Median :30778
##                     Mean   :2017   Mean   :6.266     Mean   :32929
##                     3rd Qu.:2017   3rd Qu.:7.065     3rd Qu.:41104
##                     Max.   :2017   Max.   :7.788     Max.   :94278
```

From the statistical table, it should be worth noting that:
1. There are 36 countries in the datatset;
2. The variation of the life satisfaction rates is quite low, and there seem to be no extreme values because the mean and the median are quite similar;
3. The variation of GDP across countries seems to be pretty high, and there must be some outliers because median and mean are not close to each other.

Let's verify observations 2 and 3 by looking at a **box plot**.

```
par(mfrow=c(1,2))
boxplot(GDP, main='GDP box plot',
        col = "#69c3a3",
        ylab = "GDP in $")
boxplot(Life_satisfaction, main='Life satisfaction box plot',
        col = "#fb9805",
        ylab = "Life satisfaction rate")
```





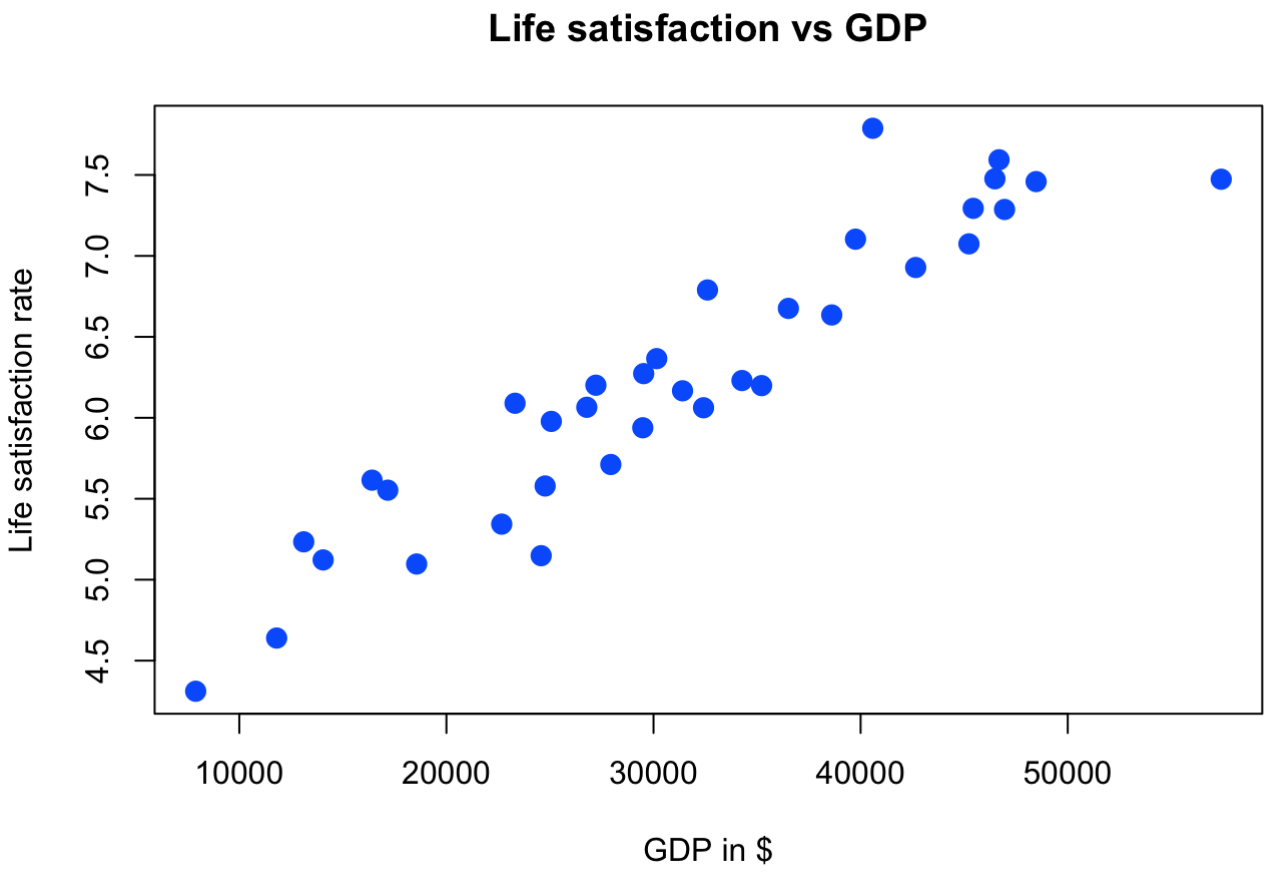
From the GDP box plot, we can observe an outlier (Luxembourg = 94277.965). In order to maintain high statistical significance of our future model, let's exclude the outlier from our dataset. We should end up with 35 countries remaining.

```
# Removing an outlier
data_new <- data[data$Country != "Luxembourg",]
nrow(data_new)
```

```
## [1] 35
```

Now, let's investigate is there any preliminary relationship between GDP and Life satisfaction. Drawing a **scatter plot** and calculating the **correlation coefficient** may help us to analyze the relationship between the variables.

```
# scatter plot
plot(data_new$GDP, data_new$Life_satisfaction,
     pch=20,
     cex=2,
     col="#0563fb",
     xlab="GDP in $", ylab="Life satisfaction rate",
     main = "Life satisfaction vs GDP")
```




```
# correlation coefficient r
print(paste("r = ", cor(data_new$GDP, data_new$Life_satisfaction)))
```

```
## [1] "r =  0.93661044460202"
```

Since our correlation coefficient is close to 1, we can conclude that there should be a positive linear relationship between GDP and Life satisfaction rate. We can also detect this relationship visually from the scatter plot. In order to check whether that's the case, we have to try to fit the linear model and perform a statistical test.

Regression Analysis

Model fitting

```
# Fitting SLR model and plotting the line
mod <- lm(data_new$Life_satisfaction ~ data_new$GDP, data=data_new)
plot(data_new$GDP, data_new$Life_satisfaction,
     pch=20,
     cex=2,
     col="#0563fb",
     xlab="GDP in $", ylab="Life satisfaction rate",
     main = "Life satisfaction vs GDP")
abline(mod, col=2, lwd=3)
```


```
# Coefficients
print(paste("B0 = ", mod$coefficients[1]))
```

```
## [1] "B0 =  4.07436713027582"
```

```
print(paste("B1 = ", mod$coefficients[2]))
```

```
## [1] "B1 =  6.95549675367128e-05"
```

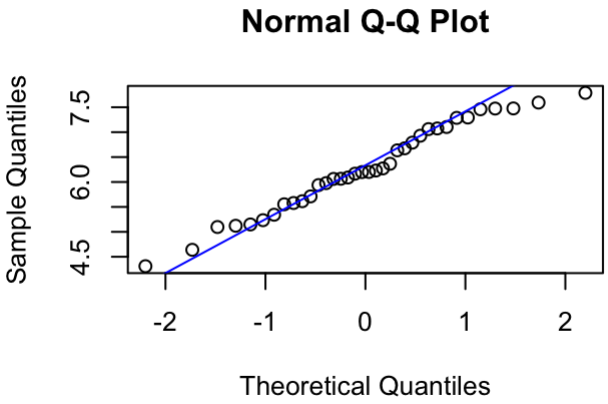
After we fitted the model, we have to check whether there is a significant linear relationship between GDP and Life satisfaction by conducting a hypothesis test concerning β_1. But before jumping into this, we first have to check whether some assumptions about our data are true.

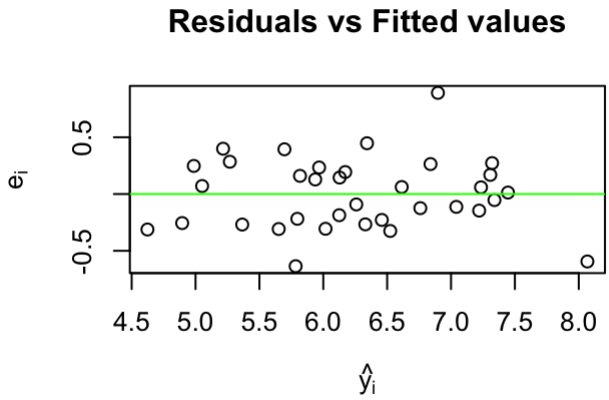
Regression Diagnostics

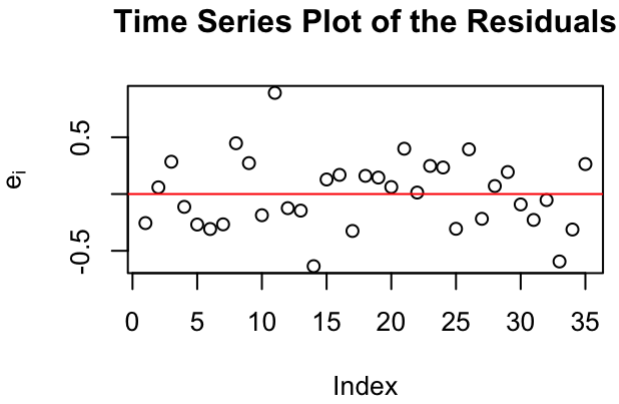
```
library(latex2exp)
par(mfrow=c(2,2))
qqnorm(Life_satisfaction)
qqline(Life_satisfaction, col = "blue")

plot(fitted(mod),mod$residuals,
     ylab=TeX("$e_i$"), xlab=TeX("\\hat{y_i}"),
     main = "Residuals vs Fitted values")
abline(h=0, col="green")

plot(mod$residuals,
     ylab=TeX("$e_i$"),
     main = "Time Series Plot of the Residuals")
abline(h=0, col="red")
```







1. **Normality**
   Our Life satisfaction rates must be approximately normally distributed. In order to check it, we will use a *Normal Q-Q plot* from above. The Q-Q line is approximately straight, which suggests that our data is approximately normally distributed.

2. **Homogeneity of variance/Linearity**
   The variance of the residuals must be constant and our linear model should be a good fit for the data. In order to check both of these assumptions, we can use *Residuals vs Fitted values* plot. Since the points in that plot are evenly spread across 0 line with no pattern, we can conclude that the variance of the residuals is constant and the linearity assumption is also satisfied.

3. **Independence**
   Our residuals also must be independent of each other. In order to check it, we can rely on *Time Series Plot of the Residuals*. From that, we can say that our residuals are independent because there is no discernible pattern in the plot.

Since all of our key assumptions are satisfied, we can proceed with hypothesis testing.

Hypothesis testing

Null hypothesis: $\beta_1 = 0$, which means there is *no* relationship between GDP and Life satisfaction.
Alternative hypothesis: $\beta_1 \neq 0$, which means there is a relationship between GDP and Life satisfaction. We will use α = 0.05

```
summary(mod)
```

```
##
## Call:
## lm(formula = data_new$Life_satisfaction ~ data_new$GDP, data = data_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63540 -0.24191  0.01309  0.21378  0.89095
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.074e+00  1.511e-01   26.96   <2e-16 ***
## data_new$GDP 6.955e-05  4.529e-06   15.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3183 on 33 degrees of freedom
## Multiple R-squared:  0.8772, Adjusted R-squared:  0.8735
## F-statistic: 235.8 on 1 and 33 DF,  p-value: < 2.2e-16
```

As we can see from the summary table, p-value for β_1 is equal to $2 \cdot 10^{-16}$, and it is smaller than our α. This means that we reject null hypothesis, and conclude that there is a linear relationship between GDP and Life satisfaction.

Conclusion

The aim of this project was to answer the question of whether the life satisfaction of people from European countries in 2017 depended on the country's GDP. The answer to this question was based on the data collected from Our World in Data (https://ourworldindata.org/grapher/gdp-vs-happiness), which is indeed a credible source because it is a scientific online publication that focuses on large global problems such as poverty, disease, inequality, etc. Its research team is based at the University of Oxford.

The investigation of the above-mentioned question started from exploratory data analysis, in which an outlier was detected. The outlier was further removed from the dataset for the sake of maintaining high statistical significance. However, we always should be very careful when removing outliers. It appears that in this case, it was too risky to remove it because the dataset was not large enough, and by removing the outlier we could lose

some important information about the data. This problem could have been solved by adding more observations into the dataset. For example, in this case, I could have added data from the previous year or two, and after that remove the outlier. In this way, it would have been much safer.

Following this, it was noticed that the correlation coefficient and the scatter plot showed that there must be a positive linear correlation between GDP and Life satisfaction. Therefore we further proceeded with testing the SLR model on our data. Before testing all the assumptions about the data were verified graphically, which may be not very precise. A better way to verify assumptions would have been to use specific tests such as Levene's or Run's test.

In the end, the hypothesis test showed that there is a linear relationship between GDP and Life satisfaction. From this, it can be concluded that there should be some amount of dependency between Life satisfaction and the country's GDP: the higher the country's GDP, the happier are the people living in this country. However, it is worth mentioning that correlation does not mean causation. There may be multiple factors affecting the Life satisfaction of people, and it would be better if in the future this fact was taken into consideration. For example, if I continued working with this project, I would definitely add other different features, such as rate of poverty, level of literacy, etc.