Gabriela Malec

Big Data and R Programming

1. Introduction

The aim of this paper is to examine variables that can have an impact on tourism. It can be achieved by analyzing the trends observed in individual countries around the world. They can be considered in relation to the economy, natural, cultural and social environment.

Hypothesis testing will be used to test this relationship, where the null hypothesis is given as:

H0 = the prosperity of countries does not affects the number of tourists arrivals

While the alternative hypothesis is:

H1= the prosperity of countries affects the number of tourists arrivals

2. Methodology

2.1. Data

The data used in this paper is mainly based on the World Economic Forum's Travel & Tourism Competitive Index dataset, supplemented by World Development Indicators from World Bank Open Data. The additional data was sourced from Our World in Data.

2.2. Variables

All variables were presented using a coherent scale so that they receive values from 1 to 7. This approach was used because when considering huge numbers such as number of tourists, population or GDP, the analysis could be less readable. And just to prevent that, the following formula was used:

$$6 * \left(\frac{\text{country value} - \text{sample minimum value}}{\text{sample maximum value} - \text{sample minimum value}}\right) + 1$$

The following indicators were adopted as independent variables:

- Population
- GDP
- HDI
- Happiness and Life Satisfaction
- Region

The Travel & Tourism Competitiveness indices were also used. They are listed below with their component variables.

- Enabling environment index
  - Business environment
  - Safety and security
  - Health and hygiene
- T&T policy and conditions index
  - Prioritization of Travel & Tourism
  - International Openness
  - Price competitiveness
  - Environmental sustainability
- Infrastructure index
  - Air transport infrastructure
  - Ground and port infrastructure
- Natural and cultural resources index
  - Natural resources
  - Cultural resources and business travel

2.3. Method

There were some missing values in the prepared dataset. Therefore, in the first step, their distribution in the dataset was analyzed and records with NA values were deleted. The following chart shows the locations of the missing data [Figure 1].

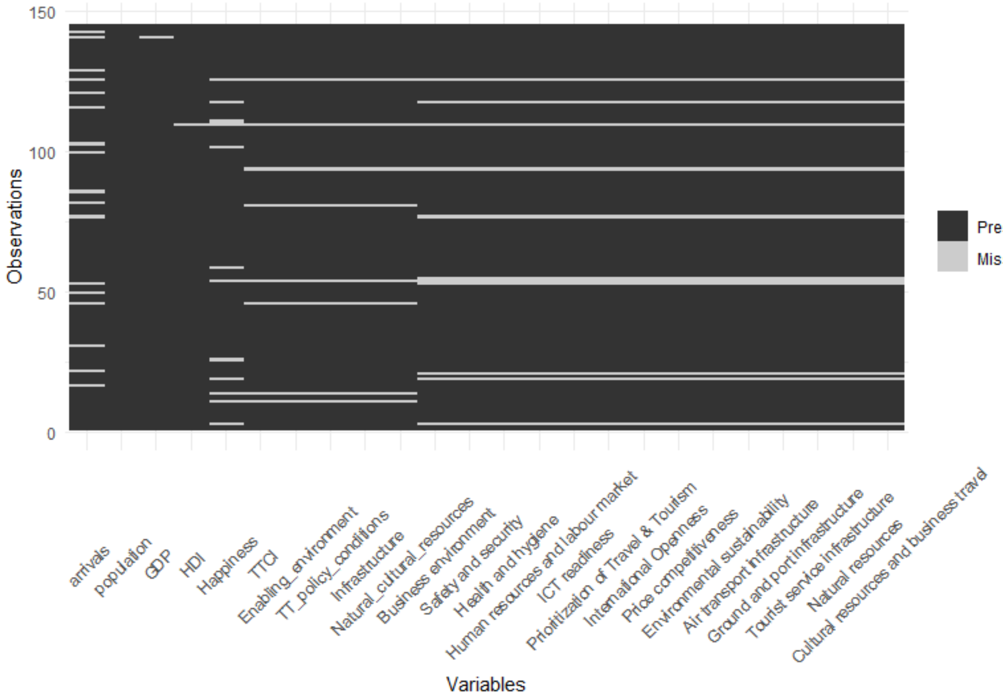


Figure 1 Chart of missing values

After that, there were 126 rows left in the dataset, representing 126 countries across 8 geographical regions. Compared to the original value of 145, it can be seen that 19 records were deleted. The regions were converted into numerical variables for analysis purposes. The boxplot [Figure 2] visualizes the distribution of tourist arrivals across different geographical regions.

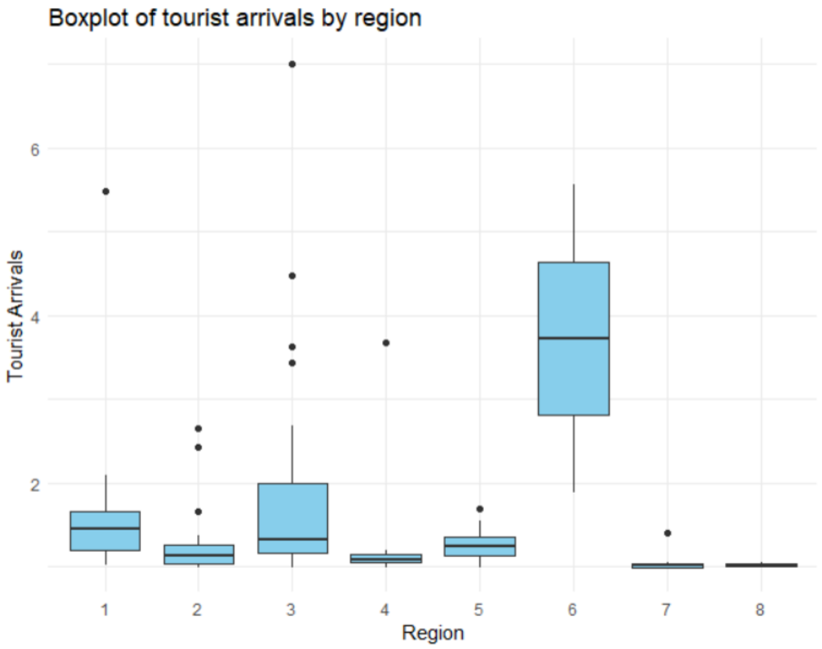


Figure 2 Boxplot

In the next step, the correlations between the variables were checked. For this purpose, a correlation chart was generated. [Figure 3] The highest correlations have been observed between the Human Development Index (HDI) and other variables, therefore this indicator will not be used in the preparation of regression models. It was assumed that the strong correlation takes the value of at least 0.75.

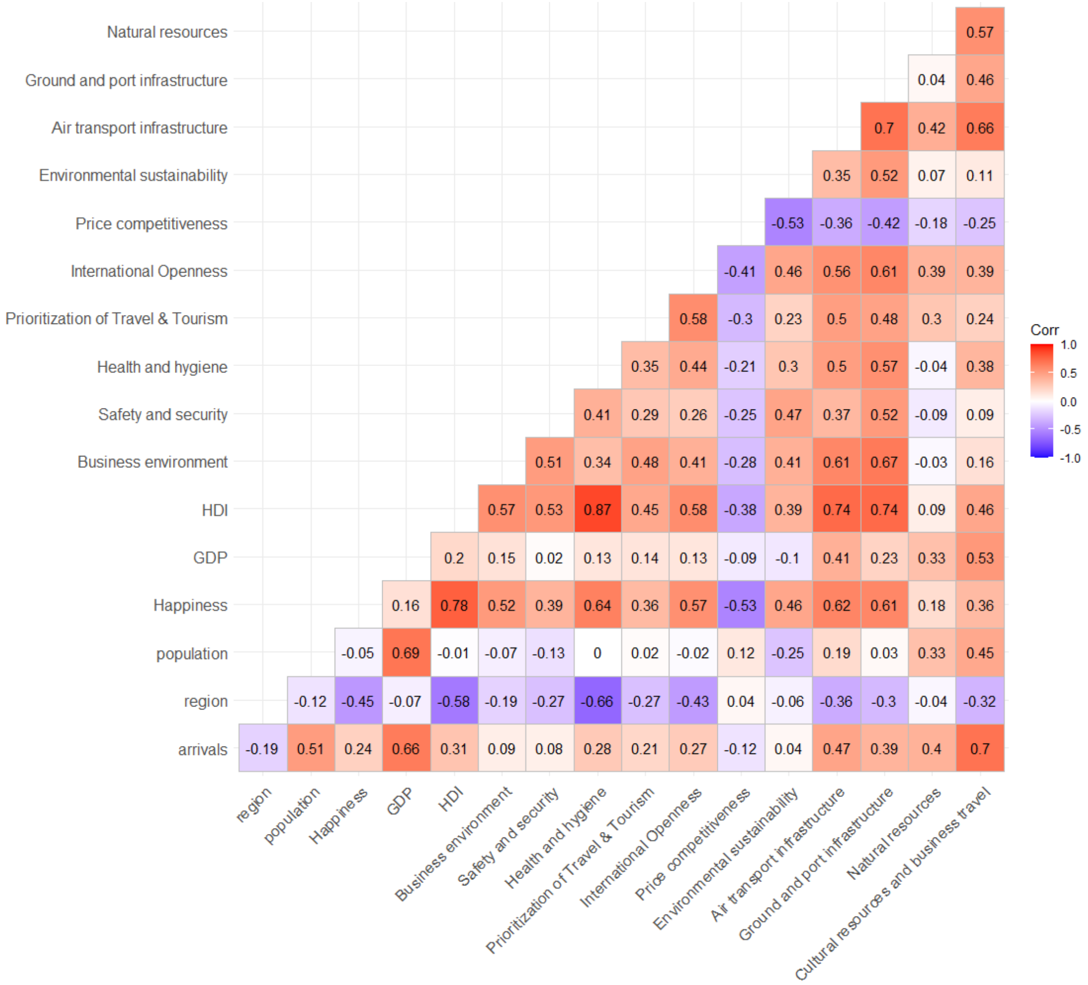


Figure 3 Correlation plot

Furthermore, a correlation matrix plot [Figure 4] was generated for a specific subset of variables. The plot provides a visual representation of the pairwise correlations among the selected variables.

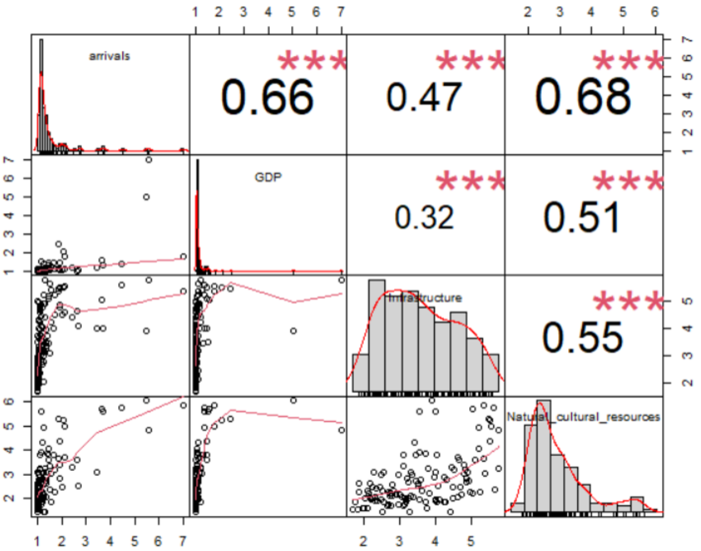


Figure 4 Correlation chart

The study focuses on a linear regression model used for the multivariable analysis to estimate the number of tourist arrivals, which is dependent variable. Indepentable variables fulfill the role of determining factors influencing the tourist attractiveness of countries. In the models, a is the intercept and βn are the different constants.

3. Results

3.1. The preliminary models

In the model 1 [M 1], the included variable was Gross domestic product (GDP). The index defines the total monetary or market value of all the finished goods and services produced within a country's borders in a specific time period. [9]

$$\textbf{Tourists} = \textbf{a} + \boldsymbol{\beta}\textbf{1} * \textbf{GDP} + \boldsymbol{\varepsilon}$$

M 1

Next model has been extended to include other variables - the region and the infrastructure development indicator [M 2].

$$\textbf{Tourists} = \textbf{a} + \beta 1 * \textbf{GDP} + \beta 2 * \textbf{Region} + \beta 3 * \textbf{Infrastructure} + \varepsilon$$

M 2

In the third model, the variable for natural and cultural resources was added [M 3].

$$\textbf{Tourists} = \textbf{a} + \beta 1 * \textbf{GDP} + \beta 2 * \textbf{Region} + \beta 3 * \textbf{Infrastructure} \\ + \beta 4 * \textbf{Natural\_cultural\_resources} + \varepsilon$$

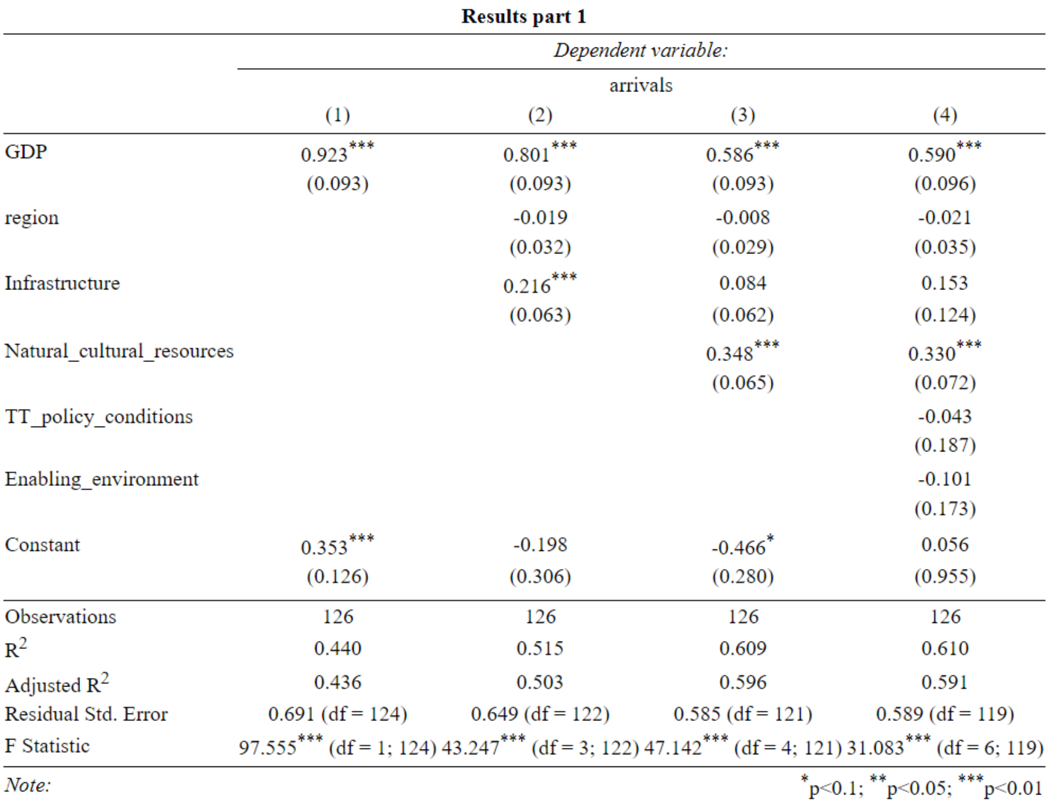
M 3

In the 4th model, 2 more variables were added. First one was T&T Policy Conditions and the second one was Enabling environment [M 4].

$$\textbf{Tourists} = \textbf{a} + \beta 1 * \textbf{GDP} + \beta 2 * \textbf{Region} + \beta 3 * \textbf{Infrastructure} + \beta 4 * \\ \textbf{Natural\_cultural\_resources} + \beta 5 * \textbf{TT\_policy\_conditions} + \beta 6 * \textbf{Enabling\_environment} \\ + \varepsilon$$

M 4

3.2. The preliminary models – summary



Results part 1

	Dependent variable:			
	arrivals			
	(1)	(2)	(3)	(4)
GDP	0.923***	0.801***	0.586***	0.590***
	(0.093)	(0.093)	(0.093)	(0.096)
region		-0.019	-0.008	-0.021
		(0.032)	(0.029)	(0.035)
Infrastructure		0.216***	0.084	0.153
		(0.063)	(0.062)	(0.124)
Natural_cultural_resources			0.348***	0.330***
			(0.065)	(0.072)
TT_policy_conditions				-0.043
				(0.187)
Enabling_environment				-0.101
				(0.173)
Constant	0.353***	-0.198	-0.466*	0.056
	(0.126)	(0.306)	(0.280)	(0.955)
Observations	126	126	126	126
R^2	0.440	0.515	0.609	0.610
Adjusted R^2	0.436	0.503	0.596	0.591
Residual Std. Error	0.691 (df = 124)	0.649 (df = 122)	0.585 (df = 121)	0.589 (df = 119)
F Statistic	97.555*** (df = 1; 124)	43.247*** (df = 3; 122)	47.142*** (df = 4; 121)	31.083*** (df = 6; 119)
Note:				*p<0.1; **p<0.05; ***p<0.01

Figure 5 Results - part 1

The results of the regression analysis are presented in the table [Figure 5]. The first fact is that in the case of all three models a high GDP index can be reflected in a large number of tourist arrivals. This is statistically significant (p<0.01). Considering other factors, it can be seen that in model 2, the level of infrastructure development may have a statistically significant impact on the dependable variable, however, after adding more variables, its significance is no longer relevant. What's more, in each of the models visible above cultural resources and business travel index stands out with statistical significance. It can be deduced that the more local and national authorities care about the development of cultural places, the more travelers choose those destinations.

Generally looking, attention should be paid to measures of the quality of model fit. To begin with, R^2 value increases with each subsequent model, and it eventually reaches a satisfactory fit. Moreover, it can be noticed that the residual std. error decreases from 0.691 in model 1 to 0.589 in model 4, so the improvement is being made by each model.

3.3. Improved models

In the next, 5th model [M 5], an attempt to improve the model was based on adding a variable concerning the population in the countries.

$$\textbf{Tourists} = \textbf{a} + \beta 1 * \textbf{GDP} + \beta 2 * \textbf{Region} + \beta 3 * \textbf{Infrastructure} + \beta 4 * \textbf{Natural\_cultural\_resources} + \beta 4 * \textbf{population} + \varepsilon$$

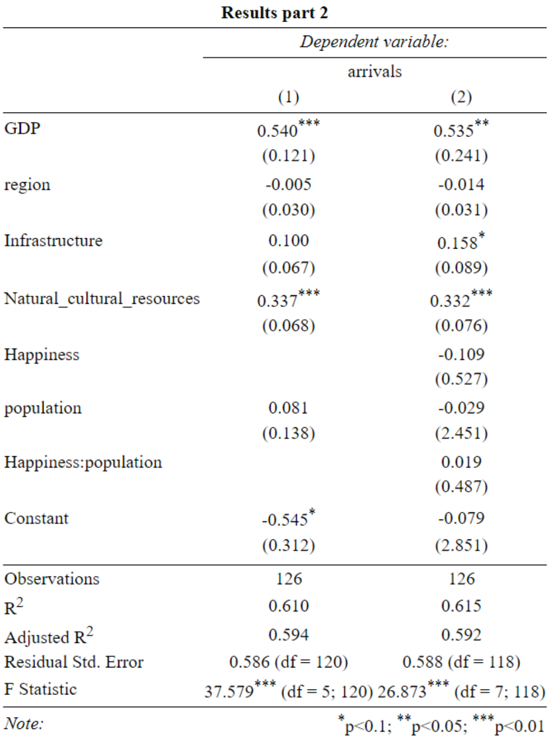
M 5

Then, in the 6th model [M 6], a variable storing information on the level of happiness and satisfaction with life by people was used, the aim was to check whether the positivity and openness of people has any real impact on the choice of the visited country.

$$\textbf{Tourists} = \textbf{a} + \beta 1 * \textbf{GDP} + \beta 2 * \textbf{Region} + \beta 3 * \textbf{Infrastructure} + \beta 4 * \textbf{Natural\_cultural\_resources} + \textbf{Happiness*population} + \varepsilon$$

M 6

3.4. Improved models - summary



Results part 2

	Dependent variable:	
	arrivals	
	(1)	(2)
GDP	0.540^{***}	0.535^{**}
	(0.121)	(0.241)
region	-0.005	-0.014
	(0.030)	(0.031)
Infrastructure	0.100	0.158^{*}
	(0.067)	(0.089)
Natural_cultural_resources	0.337^{***}	0.332^{***}
	(0.068)	(0.076)
Happiness		-0.109
		(0.527)
population	0.081	-0.029
	(0.138)	(2.451)
Happiness:population		0.019
		(0.487)
Constant	-0.545^{*}	-0.079
	(0.312)	(2.851)
Observations	126	126
R^2	0.610	0.615
Adjusted R^2	0.594	0.592
Residual Std. Error	0.586 (df = 120)	0.588 (df = 118)
F Statistic	37.579^{***} (df = 5; 120)	26.873^{***} (df = 7; 118)
Note:		*p<0.1; **p<0.05; ***p<0.01

Figure 6 Results - part 2

The results of the regression analysis of the improved models are presented in the table [Figure 6]. The attempt to achieve an improved model did not bring spectacular changes to it. Again, we can see that statistically, the high GDP index or cultural resources and business travel index, has a positive impact on the number of tourist arrivals. In the last model, we can see a rather interesting dependence that the lower the happiness index, the slightly more tourists. This is not statistically significant, but the reason for this could be, for example, that in developed countries, people tend to focus more on work, which often results in more stress and can lead to such happiness statistics. Furthermore, it can be noticed that in this model, when the higher the infrastructure index, and thus more convenient use of means of transport, the number of tourist arrival increases. It is statistically significant (p<0.1).

To further assess the fit of the models, McFadden's Pseudo-R2 was calculated for each model in both groups. The results are presented in the table below [Table 1].

McFadden					
reg1	reg2	reg3	reg4	reg5	reg6
0.2178569	0.271903	0.3526103	0.3539016	0.3536922	0.3578199

Table 1 McFadden Pseudo R2

4. Conclusion

The analysis rejects the null hypothesis that the prosperity of countries does not impact tourist arrivals. Even after introducing additional variables in model 4, the association remains highly statistically significant ($p < 0.01$), reaching ($p < 0.05$) in model 5.

In the final model, GDP exhibits a positive coefficient of 0.535, suggesting that as the GDP index, a proxy for a country's prosperity, increases, the log odds of more tourist arrivals rise by 0.577. This implies that wealthier countries tend to invest in factors influencing tourism, enhancing their attractiveness to travelers.

Although regional influence on tourist numbers wasn't observed, the study identifies positive impacts of natural & cultural resources and infrastructure development. This suggests that local authorities seeking to boost tourism may benefit from focusing on transport network development and preserving natural and cultural heritage.

Overall, each model demonstrates an improved fit, with the R-squared value increasing from 0.440 in the first model to 0.615 in the last. McFadden's R-squared also confirms the enhanced fit, rising from 0.218 in model 1 to 0.358 in model 6.

5. R script

```
# libraries

library(readxl)
library(dplyr)
library(magrittr)
library(tidyr)
library(ggplot2)
library(PerformanceAnalytics)
library(ggcorrplot)
library(stargazer)
library(scatterplot)
```

```
# upload data

tourism_data <- read_excel("tourism_data2.xlsx")

View(tourism_data2)

# simplify names of columns

names(tourism_data2) <- c('country', 'regions', 'subregions', 'arrivals', 'population','GDP', 'HDI',
'Happiness','TTCI','Enabling_environment','TT_policy_conditions','Infrastructure','Natural_cultural_re
sources', 'Business environment', 'Safety and security', 'Health and hygiene','Human resources and
labour market', 'ICT readiness', 'Prioritization of Travel & Tourism', 'International Openness','Price
competitiveness', 'Environmental sustainability', 'Air transport infrastructure', 'Ground and port
infrastructure', 'Tourist service infrastructure',  'Natural resources', 'Cultural resources and business
travel')



# check for missing values

data_num <- tourism_data2[,-c(1,2,3)]

missing_values <- function(x){

  data_num %>%

    is.na %>%

    melt %>%

    ggplot(data = .,

           aes(x = Var2,

               y = Var1)) +

    geom_raster(aes(fill = value)) +

    scale_fill_grey(name = "",

                    labels = c("Present","Missing")) +

    theme_minimal() +

    theme(axis.text.x  = element_text(angle=45, vjust=0.5)) +

    labs(x = "Variables",

         y = "Observations")

}

missing_values(df)
```

```
# remove rows with missing values
data <- tourism_data2 %>%
  drop_na()
#numbers to represent subgroups (regions)
data[data$regions == 'East Asia and the Pacific', 'region'] = 1
data[data$regions == 'Eastern Europe and Central Asia', 'region'] = 2
data[data$regions == 'Europe', 'region'] = 3
data[data$regions == 'Latin America and Caribbean', 'region'] = 4
data[data$regions == 'Middle East and North Africa', 'region'] = 5
data[data$regions == 'North America', 'region'] = 6
data[data$regions == 'Sub-Saharan Africa', 'region'] = 7
data[data$regions == 'South Asia', 'region'] = 8
# boxplot
ggplot(data, aes(x = factor(region), y = arrivals)) +
  geom_boxplot(fill = "skyblue") +
  labs(x = "Region", y = "Tourist Arrivals") +
  ggtitle("Boxplot of tourist arrivals by region") +
  theme_minimal()
# correlation
data %>%
  dplyr::select('arrivals', 'region', 'population', 'Happiness', 'GDP','HDI',
                'Business environment', 'Safety and security', 'Health and hygiene',
                'Prioritization of Travel & Tourism', 'International Openness','Price competitiveness',
'Environmental sustainability',
                'Air transport infrastructure', 'Ground and port infrastructure',
                'Natural resources', 'Cultural resources and business travel') %>%
  cor(.) %>%
  ggcorrplot(., type = "lower", lab = TRUE)
```

```
data %>%

  dplyr::select('arrivals', 'population', 'GDP', 'Happiness','Enabling_environment',
'TT_policy_conditions', 'Infrastructure', 'Natural_cultural_resources') %>%

  cor(.) %>%

  ggcorrplot(., type = "lower", lab = TRUE)


# correlation 2

data3 <- data[,c(4,6,12,13)]

chart.Correlation(data3, histogram=TRUE, pch=19)


# basic models

reg1 <- lm(arrivals ~ GDP, data=data)

summary(reg1)

reg2 <- lm(arrivals ~ GDP + region + Infrastructure, data=data)

summary(reg2)

reg3 <- lm(arrivals ~ GDP + region + Infrastructure + Natural_cultural_resources , data=data)

summary(reg3)

reg4 <- lm(arrivals ~ GDP + region + Infrastructure + Natural_cultural_resources +
TT_policy_conditions + Enabling_environment, data=data)

summary(reg4)


# improved

reg5 <- lm(arrivals ~ GDP + region + Infrastructure + Natural_cultural_resources + population,
data=data)

summary(reg5)

reg6 <- lm(arrivals ~ GDP + region + Infrastructure + Natural_cultural_resources +
Happiness*population, data=data)

summary(reg6)
```

```
# McFadden’s Pseudo R2
list(reg1 = pR2(reg1)["McFadden"],
     reg2 = pR2(reg2)["McFadden"],
     reg3 = pR2(reg3)["McFadden"],
     reg4 = pR2(reg4)["McFadden"],
     reg5 = pR2(reg5)["McFadden"],
     reg6 = pR2(reg6)["McFadden"])


# print table with model info in html code
stargazer(reg1, reg2, reg3, reg4, title="Results part 1", type='html',align=TRUE)
stargazer(reg5, reg6, title="Results part 2", type='html',align=TRUE)


# plot
lm_fit <- lm(arrivals ~ GDP + region + Infrastructure + Natural_cultural_resources + Happiness*population, data=data)
summary(lm_fit)
predicted_df <- data.frame(mpg_pred = predict(lm_fit, data), arrivals=data$arrivals)
ggplot(data = data, aes(x = GDP, y = arrivals)) +
    geom_point(color='blue') +
    geom_line(color='red',data = predicted_df, aes(x=mpg_pred, y=data$arrivals))
```