BUSINESS INTELLIGENCE

Final report

Table of Contents

Executive summary

It is important to measure the effects of investments in research and innovation. Four critical success factors are analyzed: input, output, quality of the output, and reputation of the output. Together, these four critical success factors (CSFs) provide a cohesive view on to what degree the investments from the European Commission and the European countries have led to the desired results.

The CSFs are measures by a few key performance indicators (KPIs) each, and the data gathered to measure these KPIs come from reliable sources either from the EU themselves or from sources audited by professional firms. The data has been cleaned and transformed such that the statistical meaning is the same but the quality of the results derived thereof are improved. The data are stored in a database that is designed with a star schema, meaning that the data can be viewed in two dimensions: geographical area and time.

A challenging crucial domain for the whole project is data analytics. In this chapter the research focuses on the investigation of the manually collected data set. The first step is to identify and get a hold on the structure of the dataset as well as to discover the measures of central tendency: mean, median, mode, quartiles, and the measures of dispersion/variation: standard deviation, variance, and range. Since the descriptive statistics have been identified and reported, the research focus on the creation of predictive models that will provide information about the future academic performance of the countries under investigation.

Dashboards are of crucial importance in today's age for executives to get a grasp on what is happening in their company or in the world. They aim to consolidate data into smart visualisations that can be interacted with. The better the interaction possibilities with data, the more intuitive the design and the more powerful functionalities of a dashboard can have tremendous positive impact on management decisions. Therefore, we have created a dashboard with the following four key principles in mind: clarity, intuitiveness, usefulness and interactivity. The dashboard is divided into five 'tabs' where four tabs represents a critical success factor which each houses multiple key performance indicators and the fifth tab shows an overview of the performance of each country for the most recent year.

Based on the dashboard, several recommendation can be made for each country. Austria is performing poorly all across the board and should reassess the effectiveness for its large investments in public R&D expenditure and education expenditure. Germany performs decently overall, but for each CSFs it performs poorly on one KPI. Germany should thus analyse the composition of the R&D expenditure to determine if enough of it is going towards improving the KPIs on which Germany is performing poorly. The Netherlands is performing excellent all across the board and should focus on maintain their currently quality and perhaps do additional investments in its universities.

The following improvements will be made to the dashboard given additional resources: add more countries, allow for comparisons between countries on multiple geographical dimensions, implement a more statistical calculation of the performance thresholds for the KPIs on the university dimension, allow the user to insert their own performance thresholds, and further evaluate and validate whether the current set of KPIs are sufficient and correct to measure their respective CSFs.

Two notebooks can be found on git: one with the dashboard and one with the code for the analytics from Data analytics chapter. The dashboard .rmd is named "final dashboard" and the data analytics code is called "Data analytics chapter".

BI strategy

The European Commission (EC) has for years stimulated investment in European research, so that more innovation and high quality jobs are created in Europe (http://ec.europa.eu/invest-in-research/index_en.htm). For the upcoming budget plans, i.e. the budget plan for the years 2021-2027, the intention is to double the budget for the EU research and innovation programme (https://ec.europa.eu/research/index.cfm?pg=newsalert&year=2017&na=na-030717). The effects of such investments must be measured in order to account for such investments, particularly if the EC intents to double the budget. As such, this document will detail a business intelligence solution that allows the EC to measure the effects of its investments in research and innovation in different European countries, and even to compare these countries with each other.

The business intelligence solution in this document regards an initial prototype, in the sense that all functionality is present but only for a subset of countries. The countries that can be analysed with this application are: the Netherlands, Germany, and Austria. In terms of development in innovation capability, these three countries allow for interesting comparisons to be made: the Netherlands is both one of the biggest innovators and one of the fastest growing innovators, Germany is one of the biggest innovators, and Austria is one of the fastest growing innovators (http://ec.europa.eu/growth/industry/innovation/facts-figures/scoreboards/).

Below is discussed in detail how the effects of investments in research and innovation is measured. This is done by first identifying critical success factors (CSF), which can be seen as elements that lead to success. Then, for each critical success factor, several key performance indicators (KPI) are chosen. A KPI provides a quantitative way to measure the degree to which a CSF is achieved. In total, four CSFs have been identified.

Critical success factor #1 – Input in research and innovation

Growth cannot result from taking no action, it has to be stimulated in some way. As such, the way countries stimulate growth in research and innovation is an important factor to look at, as it is only when a country is making the proper investments that a country can see progress.

For this critical success factor three KPIs are measured:

- **Public research & development expenditure**: this KPI is intended to track how much the government and higher education invest in research & development (R&D), and is calculated as percentage of the gross domestic product (GDP). It is the most important KPI, as more investments in R&D result in more research and innovation projects, such as research grants.
- **Human resources**: ultimately it’s the scientists and engineers that are in charge of transforming the investments into new research and innovation. This KPI is intended to measure the scientists and engineers of a country as percentage of the active population. The idea is that a higher percentage of scientists and engineers of the active population will lead to more research and innovation, as a greater part of the country is working in the academic knowledge sector.
- **Education expenditure**: more investments in education could make education more accessible to a larger part of the population or increase the quality of the education. As such, an increase in education expenditure could either result in more potential scientists and thus in more research and innovation, or in better scientists and thus better research and innovation. Education expenditure is calculated as percentage of the GDP.

Critical success factor #2 – Output of the investments in research and innovation

If a country is investing in research and innovation, then there should be a noticeable effect in that country’s amount of research and innovation. If there is not such an increase in output, the country should analyse its investments efforts to ensure growth.

For this critical success factor three KPIs are measured:

- **Citations**: a citation occurs when one scientists refer to the work of another scientist. In other words, if a paper gets cited, it means that knowledge from the cited paper gets spread. It is also an indication of acclaim: the knowledge of the research does not just exist, it gets adopted.
- **Highly cited papers**: this KPI is an extension of the previous KPI and measures the share of publications amongst the top 10% cited publications of a year. This KPI indicates a country's contribution to the most adopted knowledge.
- **Number of patent applications**: a patent provides exclusive rights to the result of innovation, and thus is a good indicator of how much innovation is created in a country. The patent applications are measured per million of the active population.

Critical success factor #3 – Quality of the output

Growth in itself is not necessarily desirable: it has to be of at least the same quality, or preferably of an even higher quality. If growth is accompanied by a decrease in quality, then the benefits of research and innovation to society (e.g. higher quality of life, more jobs) are diminished. The quality of the output is therefore of critical importance to the success of research and innovation investments.

For this critical success factor four KPIs are measured:

- **Quality of teaching**: the teaching that students and future scientists undergo develops them, their interests, and their quality of thinking. Investments in research and innovation should thus also result in a higher quality of teaching.
- **Research**: this KPI measures the quantity of the research but also the quality thereof. An increase in the amount of research and innovation is more meaningful if the quality thereof also increases.
- **Innovation score**: this KPI is a score measured by the EC themselves, and is an assessment of the innovation performance of European countries. A good innovation score indicates that a country is not only innovating a lot, but also in a healthy variety of areas (e.g. finance, firm investments).
- **Value added**: measures the value added of knowledge intensive services as percentage of total value added. Research and innovation of higher quality should result in a noticeable effect in the value added, and thereby the economy.

Critical success factor #4 – Reputation of the country's research

Not only the quality of the output is of importance, but also the international reputation thereof. If a country's research is of high reputation, then it gets adopted by researchers of other countries. Furthermore, knowledge from a country might become part of standard or global frameworks.

For this critical success factor four KPIs are measured:

- **International outlook**: this KPI measures the ability of a university to attract students and staff from other countries, as well as collaboration with universities from foreign countries. A good international outlook indicates that the quality of a university and their research is acknowledged by the international community.
- **Industry income**: this KPI measures the ability of a university to attract funding from businesses to perform research. If a business funds research from a university, then that it means that the business sees value in the university's research and innovation capabilities.

Data sources and ETL

This section is intended to describe the journey of the data from source to dashboard. First, the sources of the data and their trustworthiness are discussed. Then, extraction, transformation, and loading of this data is described. Finally, examples of R code are used to show what happens to the data from when it is queried from the data warehouse to being plotted in the dashboard.

Data sources

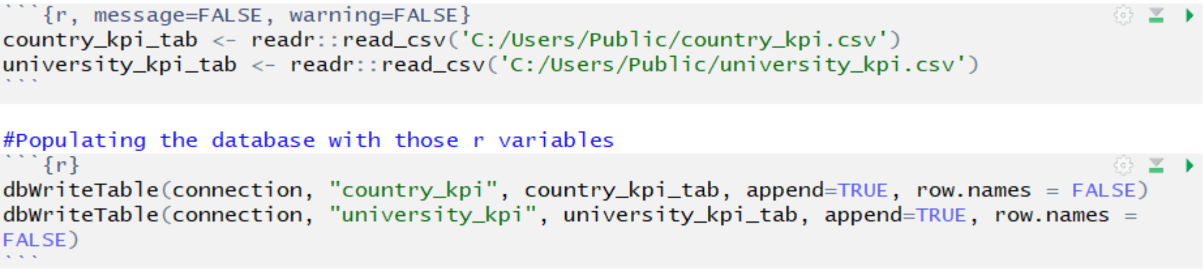
The list of data sources can be found in Appendix A – Data sources. A total of four sources have been used: Times Higher Education, Eurostat, the Research and innovation European Commission website, and the European Innovation Scoreboard. Data from Eurostat, the Research and innovation European Commission website, and the European Innovation Scoreboard are owned and governed by the European Union, and are therefore considered a reliable source to measure the KPIs. Times Higher Education is independent from the European Union, but is globally considered to be amongst the best sources for university rankings. Furthermore, the methodology used by Times Higher Education is transparent (https://www.timeshighereducation.com/world-university-rankings/methodology-world-university-rankings-2018) and has been audited by PricewatehouseCoopers, one of the four biggest Auditing firms of the world. As such, data from Times Higher Education can also be considered as reliable and trustworthy.

ETL

The data was extracted manually from the sources into an Excel file. For data from Times Higher Education, the list of universities was filtered on country, after which for each year all data for each university from a country was copied. For data from Eurostat and the Research and innovation European Commission website, the data explorer on that website was used to get a table of the required data for each country for all years, after which the relevant data was copied. For data from the European Innovation Scoreboard, the individual reports on each country under analysis was downloaded, from which the required data was taken.

Because data was already aggregated at the required level, not much action was undertaken during the data transformation step. The data was formed as a table matching the data warehouse design (see Data warehouse). Furthermore, country codes (e.g. GER) were transformed to the complete country name (e.g. Germany). Additionally, some special signs such as ‘ë’ were simplified to ‘e’ to allow for easy loading into the data warehouse. Finally, the file was saved as a CSV file.

The data in the CSV files was loaded in the data warehouse using the programming language R. The R code used can be found in Figure 1. In the top two lines of code the CSV files are saved as a data frame using the read_csv function. In the bottom two lines of code the data frames are written to the data warehouse.



~~~
```{r, message=FALSE, warning=FALSE}
country_kpi_tab <- readr::read_csv('C:/Users/Public/country_kpi.csv')
university_kpi_tab <- readr::read_csv('C:/Users/Public/university_kpi.csv')
```

#Populating the database with those r variables
```{r}
dbWriteTable(connection, "country_kpi", country_kpi_tab, append=TRUE, row.names = FALSE)
dbWriteTable(connection, "university_kpi", university_kpi_tab, append=TRUE, row.names =
FALSE)
```
~~~

Figure 1 – Data loading during the ETL process

Dataflow

The dataflow diagram can be found in figure 2.

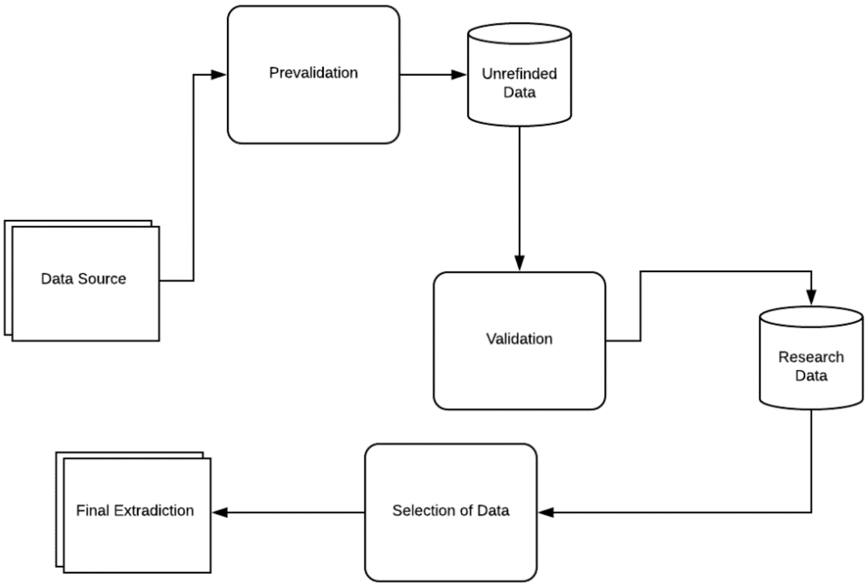


Figure 2 – Dataflow diagram

Pre- validation: The process of documenting the available data from the data sources and understanding the wider solution that they provide.
Unrefined Data: After the pre- validation the unrefined data are stored in a database.
Validation: Making sure that the data are in the best possible form.

Data warehouse

In this section the design and implementation of the data warehouse is discussed. First, a rationale is given for the data warehouse tables and for saving data in this particular way. Then, the actual data warehouse implementation is discussed along with examples of R code.

Dimensional model and explanation

In Figure 3 the designed dimensional model of the data warehouse can be found. A star schema design has been followed, meaning that the data warehouse is designed so that there is a fact table that can be seen from multiple dimensions. For this data warehouse, two fact tables were made. This was necessary because some KPIs could only be measured for universities, whereas other KPIs could only be measured for countries:

- KPIs for countries: public research& development expenditure, human resource, education expenditure, highly cited papers, number of patent applications, innovation score, and value added
- KPIs for universities: citations, research, international outlook, and industry income

A total of three dimensions can be used to view the data. The first dimension is time, which indicates to which time specification a fact relates. For example, a fact can represent data for an entire year, but also for a quarter. The second dimension is country, which indicates to which geographical specification a fact relates. For example, a fact can relate to an entire country, but also to a province or a city. The third dimension is university, which indicates to which educational institution a fact relates. For example, a fact can be about a university or some secondary institution.

A variety of primary and foreign keys are used to make the data warehouse relational. Most notably, the primary keys of the dimension tables are used as foreign keys in the fact tables.

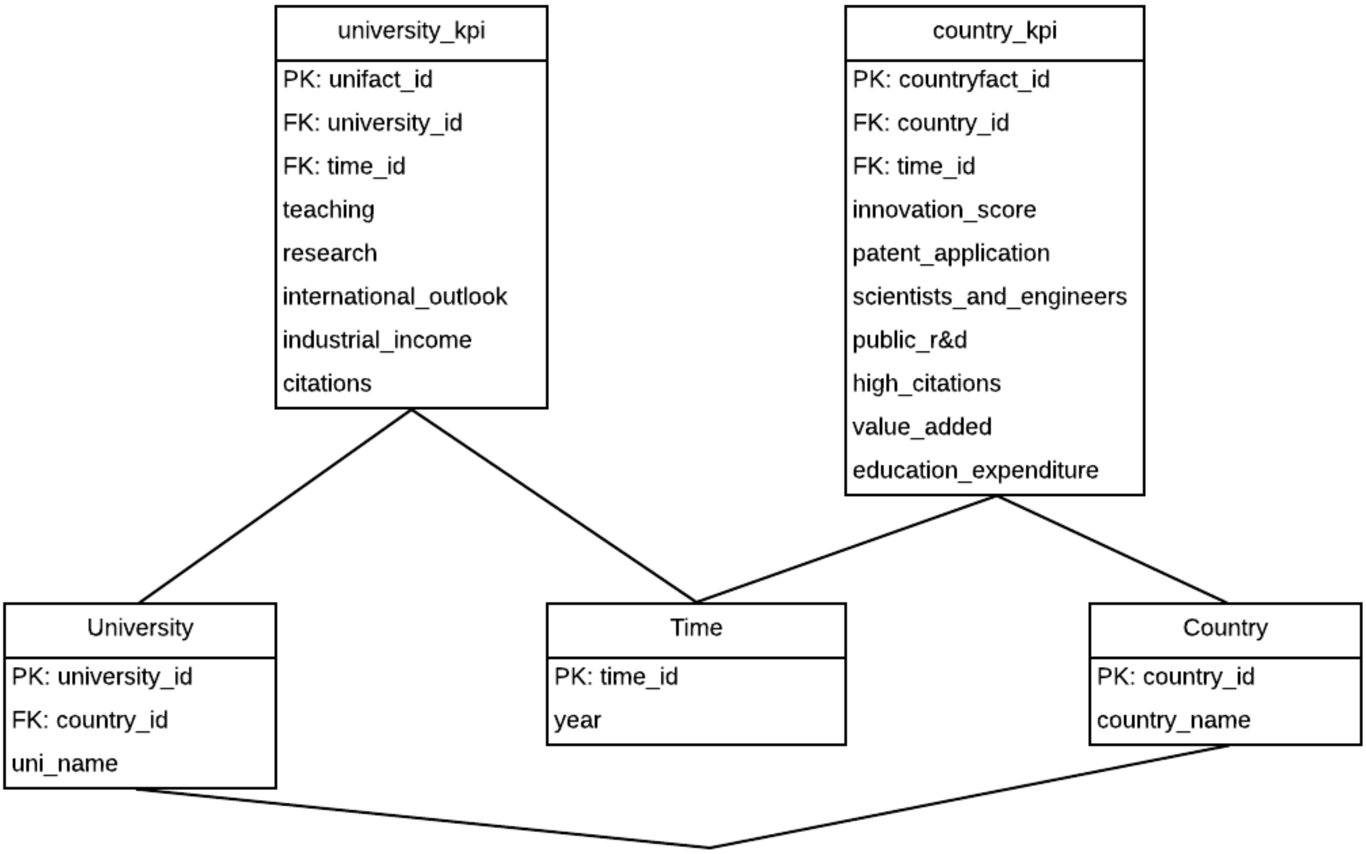
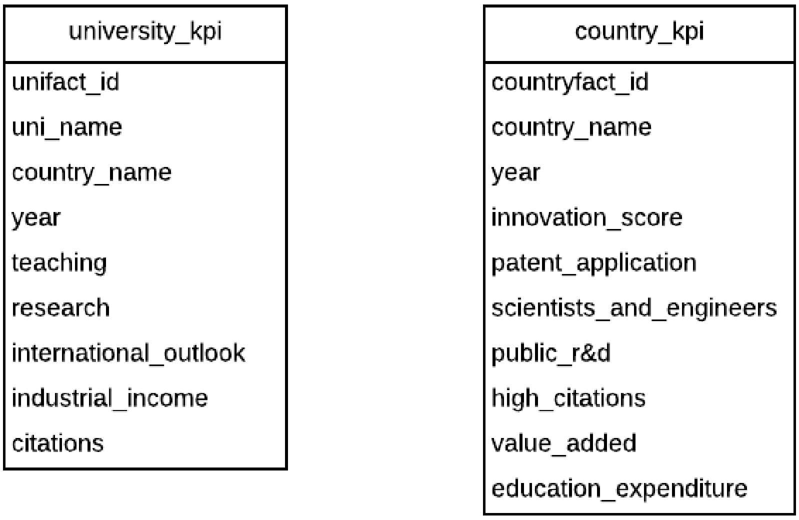


Figure 3 – Designed dimensional model of data warehouse

Actual implementation

In Figure 4 the actual dimensional model of the data warehouse can be found. This dimensional model of the actual data warehouse is notably simpler, because there are no longer any dimension tables and the data warehouse is no longer relational. The philosophy of the designed data warehouse is still important because it gives an indication of what the important dimensions are and how the data can be analysed. However, it proved to be unnecessarily complex, because the dimensional tables would consist of only one entry. For the dimensional table Time, only the year is used as a dimension. For the dimensional table Country, only the country name is used as an dimension. And finally, for the dimensional table University, only the university is used as a dimension. Because the dimensional tables were removed, so too was the need for primary and foreign keys. In the actual implemented model "country_name" could be a primary key in one table and a foreign key in another, but this was once again unnecessarily complex.

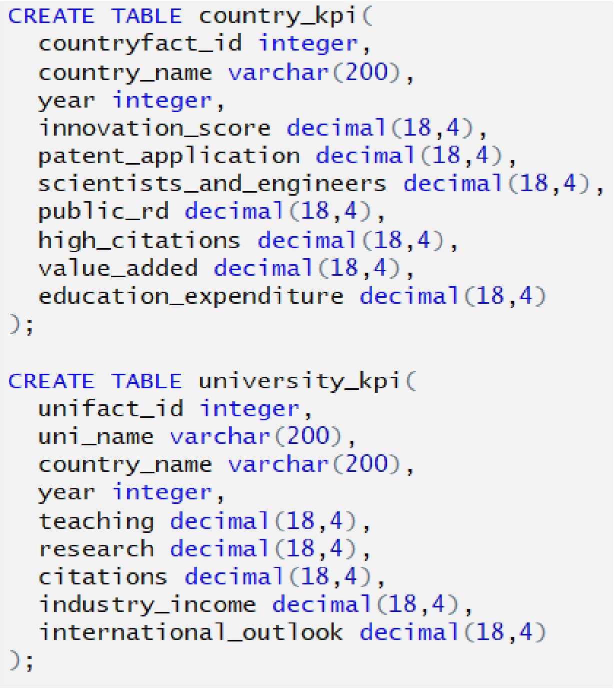


university_kpi
unifact_id
uni_name
country_name
year
teaching
research
international_outlook
industrial_income
citations

country_kpi
countryfact_id
country_name
year
innovation_score
patent_application
scientists_and_engineers
public_r&d
high_citations
value_added
education_expenditure

Figure 4 – Actual dimensional model of the data warehouse

The actual dimensional model of the data warehouse was implemented using R, for which the code can be found in Figure 5. The data types used are integer for integer numbers (such as the year), decimal for numbers with a decimal (all KPI variables), and varchar for text-based data (such as the country name).



```
CREATE TABLE country_kpi(
  countryfact_id integer,
  country_name varchar(200),
  year integer,
  innovation_score decimal(18,4),
  patent_application decimal(18,4),
  scientists_and_engineers decimal(18,4),
  public_rd decimal(18,4),
  high_citations decimal(18,4),
  value_added decimal(18,4),
  education_expenditure decimal(18,4)
);

CREATE TABLE university_kpi(
  unifact_id integer,
  uni_name varchar(200),
  country_name varchar(200),
  year integer,
  teaching decimal(18,4),
  research decimal(18,4),
  citations decimal(18,4),
  industry_income decimal(18,4),
  international_outlook decimal(18,4)
);
```

Figure 5 – R code of data warehouse creation

Architecture

During the ETL process, only Excel was used to create and transform the CSV dataset to be loaded into the data warehouse. Postgresql was used for the creation of the data warehouse and for connecting to this data warehouse. The R package readr was used so that the CSV files could be loaded into R. The R package dplyr was used to manipulate the data, for example to select a certain variable or to filter on variable values. The R packages ggplot2 and GGally were used to provide some extra functionality to the plotting function of R (ggplot). Finally, the R packages shiny and shinydashboard were used to build the dashboard.

Data analytics

Descriptive: The use of data to find out what happened in the past

The countries data frame contains 1458 observations and 10 variables that describe the performance of 3 countries (Austria, Netherlands and Germany), in research and innovation. The variables are: countryfact_id integer type, country_name character type, year numerical type ranging from 2000 to 2017, innovation_score numeric type ranging from 113 to 130 with 891 Na's , patent_application numeric type ranging from 347,5 to 588,9 with 648 Na's, scientists_and_engineers numeric type ranging from 2,1 to 9,2 , public rd numeric type ranging from 0,64 to 0,94 with 162 Na's, high citations numeric type ranging from 9,10 to 14,90 with 243 Na's, value-added numeric type ranging from 31,97 to 47,79 with 81 Na's and education expenditure ranging from 4,430 to 5,980 with 189 Na's.
The universities data frame contains 14256 observations and 9 variables that describe the performance of 528 universities from 3 countries (Austria, Netherlands, Germany), in research and innovation. The variables are: unifact_id integer type, uni_name character type, country_name character type, year integer type ranging from 2011 to 2018, teaching numeric type ranging from 9 to 70,50 with 3969 Na's, research numeric type ranging from 11,10 to 77,40 with 3969 Na's, citations numeric type ranging from 21,70 to 96,70 with 3969 Na's, industry_income ranging from 9,10 to 100 with 4482 Na's and internationals_outlook numeric type ranging from 24,30 to 100 with 3996 Na's.

Measures of central tendency

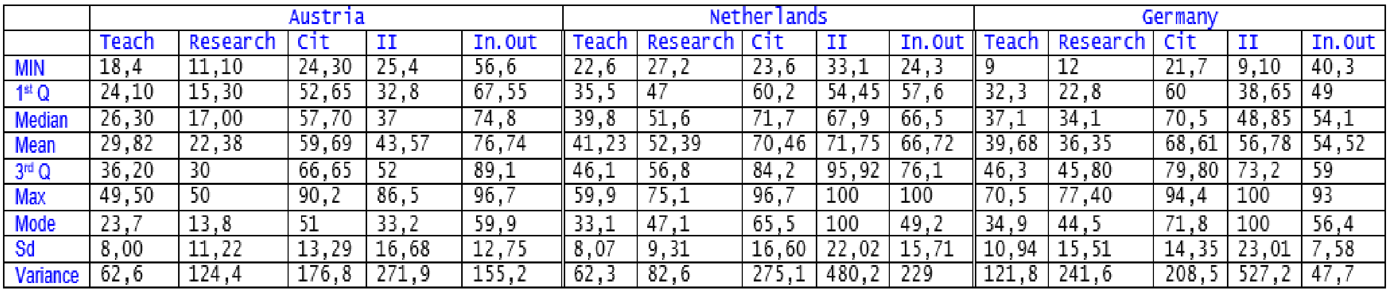
KPIs->innovation_score, patent_application, scientists_and_engineers, public_rd, hgh_citations, value_added, education_expenditure

	Austria							Netherlands							Germany						
	IS	PA	S&E	PR	HC	VA	EE	IS	PA	S&E	PR	HC	VA	EE	IS	PA	S&E	PR	HC	VA	EE
MIN	112	419,5	3,50	0,68	10,8	32,63	5,4	119	347,5	5,9	0,83	14,3	47,04	5,53	124	494,6	5,8	0,89	11,50	34,25	4,63
1st Q	113	424	4,9	0,70	11,1	32,82	5,49	119	377,1	7,9	0,84	14,5	47,58	5,590	127	513,7	6,9	0,91	11,55	35,24	4,65
Median	114	433,9	5,3	0,85	11,3	32,9	5,62	127	383,5	8	0,87	14,7	47,66	5,890	129	530,4	7	0,92	11,50	35,74	4,68
Mean	114,1	433,8	5	0,79	11,3	32,85	5,6	124	378,7	7,76	0,86	14,62	47,56	5,784	127,6	532,1	6,8	0,91	11,6	35,51	4,804
3rd Q	115	441,3	5,6	0,85	11,5	32,94	5,8	128	390,1	8,4	0,88	14,7	47,76	5,930	129	557,5	7,1	0,916	11,5	35,83	4,98
Max	121	450,1	5,7	0,87	11,8	32,95	5,91	129	395,3	8,6	0,9	14,90	47,79	5,98	129	564,1	7,2	0,93	11,60	36,40	5,08
Mode	115	383,77	2,1	0,68	9,1	31,9	5,4	119	347,49	5,3	0,83	13,5	41,33	5,5	129	494,62	5,3	0,74	11,5	34,24	4,57
Sd	1,19	11,22	0,8	0,08	0,34	0,11	0,19	4,3	16,9	0,9	0,02	0,2	0,27	0,18	3,84	693,93	0,26	0,0002	0,04	0,28	0,18
Variance	3,84	693,93	0,26	0,002	0,002	0,08	0,03	19,2	285,66	0,94	0,0006	0,04	0,07	0,03	3,84	693,93	0,26	0,002	0,002	0,08	0,03

Figure 6 Measures of Central Tendency 1

Mean: For Innovation Score variable Germany and Netherlands appear to have the same average score, whereas Austria has a smaller mean from the two countries. Next is the variable of Patent Application variable where Germany is the leader in this section meaning that Germany focus on creating new patents, next comes Austria and the last position are the Netherlands. Following with the Scientists and Engineers variable the Netherlands lead with the highest mean, implying that there are considerably higher educated workers in the Netherlands, second place goes to Germany and the smallest mean is achieved by Austria. Next variable Public Rd the Netherlands and Germany appear to have almost the same mean, whereas Austria is following. Following with High Citations the Netherlands lead the section, meaning that the Netherlands provide quality material for citations, Germany and Austria are following with almost the same average in High Citations. Next variable is Value Added the Netherlands lead with the highest mean, implying that the academics create additional value for the society, Germany and Austria appear to have almost the same average in additional value. Education Expenditure Austria and the Netherlands appear to have the same average score, therefore both countries are investing for quality in education. Germany follows the two countries but with a smaller mean.

KPIs->innovation_score, patent_application, scientists_and_engineers, public_rd, hgh_citations, value_added, education_expenditure



	Austria					Netherlands					Germany				
	Teach	Research	Cit	II	In.Out	Teach	Research	Cit	II	In.Out	Teach	Research	Cit	II	In.Out
MIN	18,4	11,10	24,30	25,4	56,6	22,6	27,2	23,6	33,1	24,3	9	12	21,7	9,10	40,3
1st Q	24,10	15,30	52,65	32,8	67,55	35,5	47	60,2	54,45	57,6	32,3	22,8	60	38,65	49
Median	26,30	17,00	57,70	37	74,8	39,8	51,6	71,7	67,9	66,5	37,1	34,1	70,5	48,85	54,1
Mean	29,82	22,38	59,69	43,57	76,74	41,23	52,39	70,46	71,75	66,72	39,68	36,35	68,61	56,78	54,52
3rd Q	36,20	30	66,65	52	89,1	46,1	56,8	84,2	95,92	76,1	46,3	45,80	79,80	73,2	59
Max	49,50	50	90,2	86,5	96,7	59,9	75,1	96,7	100	100	70,5	77,40	94,4	100	93
Mode	23,7	13,8	51	33,2	59,9	33,1	47,1	65,5	100	49,2	34,9	44,5	71,8	100	56,4
Sd	8,00	11,22	13,29	16,68	12,75	8,07	9,31	16,60	22,02	15,71	10,94	15,51	14,35	23,01	7,58
Variance	62,6	124,4	176,8	271,9	155,2	62,3	82,6	275,1	480,2	229	121,8	241,6	208,5	527,2	47,7

Figure 7 Measures of Central Tendency 2

For the evaluation of the Universities the following variables were utilized: Teaching, Research, Citations, Industry Income and International Outlook. Except International Outlook where Austria is by far the leader, the Netherlands is always leading meaning that the academic quality is high.

Predictive: The use of data to find out what could happen in the future

Prescriptive—the use of data to prescribe the best course of action for the future. A predictive model has the specific objective of allowing us to predict the value of some target characteristic of an object on the basis of observed values of other characteristics of the object.

Since the data are loaded in R, transforming the variables is the next goal. It is necessary to create data frames, to execute operations on the variables. The dataset is called countries, as a first step is to omit the NA values from the dataset, also it is important to remove the columns that are not necessary e.g. country_fact_id. The next step is to investigate if normality applies to the dataset. Through an inspection of the histograms the variables appear not to be normally distributed, therefore the transformation of data is necessary to create a predictive model.

We can see from the model output that both girth and height are significantly related to volume, and that the model fits our data well.

For the multiple linear regression, the predictor variables are: education_expenditure, patent_application, scientists_and_engineers, public_rd, high_citations and value_added. The response variable is going to be innovation_score.

The multi linear regression showed that the model accounted for 90% of the variance and all the investigated variables are significant for the explanatory variable.

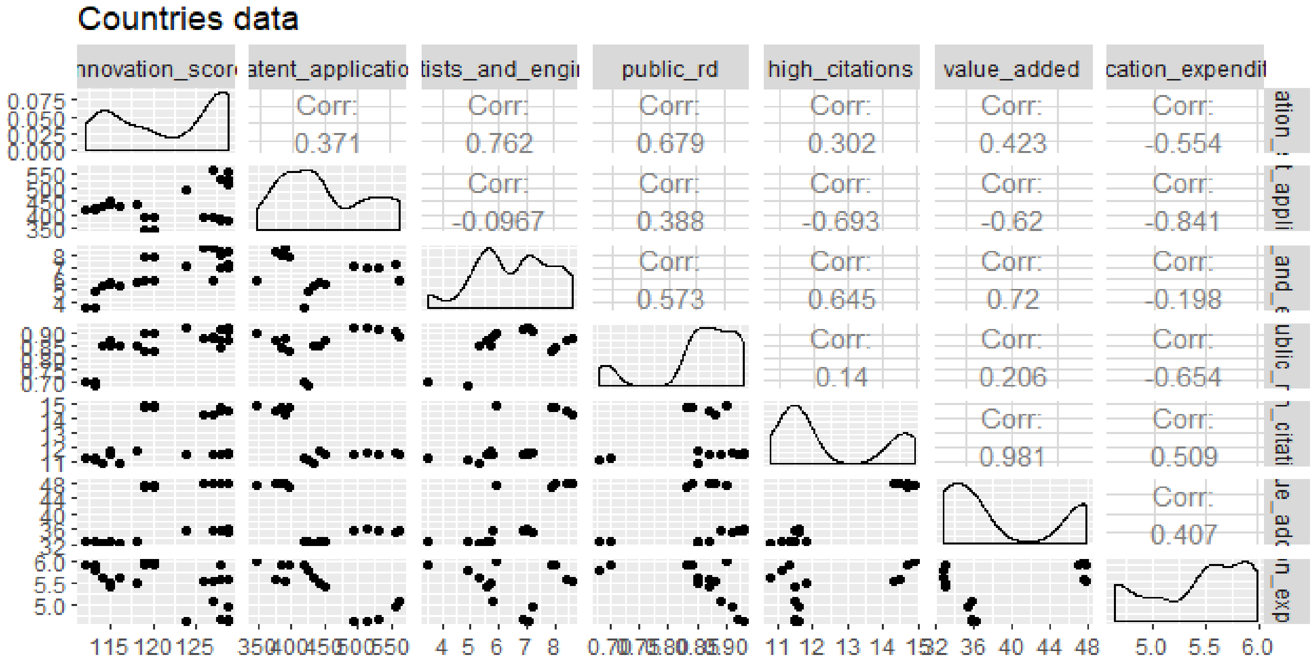


Figure 8 Variable Correlation 1

The matrix plot above visualizes the relationship among all variables in one single image. Fitting a linear regression on this dataset and see how well it models the observed data. Predictors will be added to give each of them a separate slope coefficient.
For our multiple linear regression example, we want to solve the following equation:
Innovation_score=B0+B1∗ education_expenditure +B2∗ patent_application +B3∗ scientists_and_engineers + B4* public rd + B5* high_citations +B6* value_added
The model will estimate the value of the intercept (B0) and each predictor's slope (B1) for education_expenditure, (B2) for patent_application, (B3) scientists_and_engineers, (B4) for public rd, (B5) high_citations, (B6) value_added. The intercept is the expected innovation_score for the value across all predictors. The value for each slope estimate is the increase in innovation_score related with a one-unit increase in each predictor value, holding the others constant. For better prediction purposes, the utilization of two subsets is necessary. The first is called trainSize with 75% of the observations and the other is called testSet with the remaining 25% of the observations.

The results of the models: Estimated innovation_score is 112 in comparison to the test model which is 78.
For improving one percentage point of education_expenditure there will be a decline in innovation_score by 6,65% in the training model, close to the2,7% decline in the test model. If there is an improvement in value_added by one point, it will lead for the training model to an extra 1,3 % for innovation_score and 1,1% increase for the test model in the test model. Note that R-squared number of 0.9 (compared to 0.8 from the test model). Recalling the results from the two models, both had a significant p-value (close to zero). But the difference between the two models is that scientists_and_engineers, public_rd and high_citations are not significant for the test model. However, the most important fact is that F-statistic dropped from 259,7 in the training model to 72,87 in the test model explaining the fact that there are more outliers in the later. (Results Are available in Appendix C). Another important fact is that the p-value is less than 0,05(p-value: < 2.2e-16) meaning that there is significant relationship between the variables in the dataset.

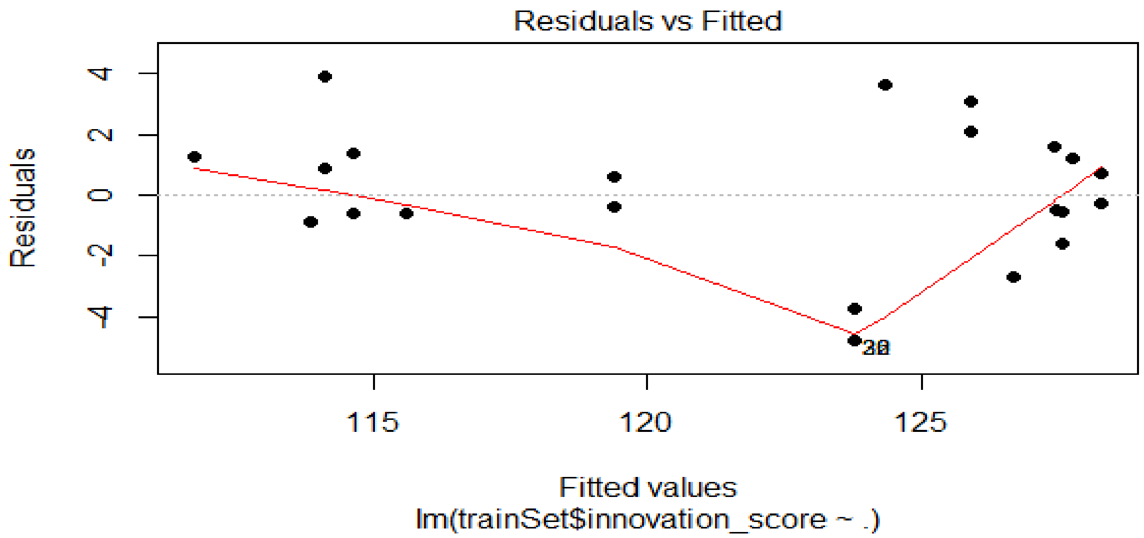


Figure 9 TrainSet Regression Outcome

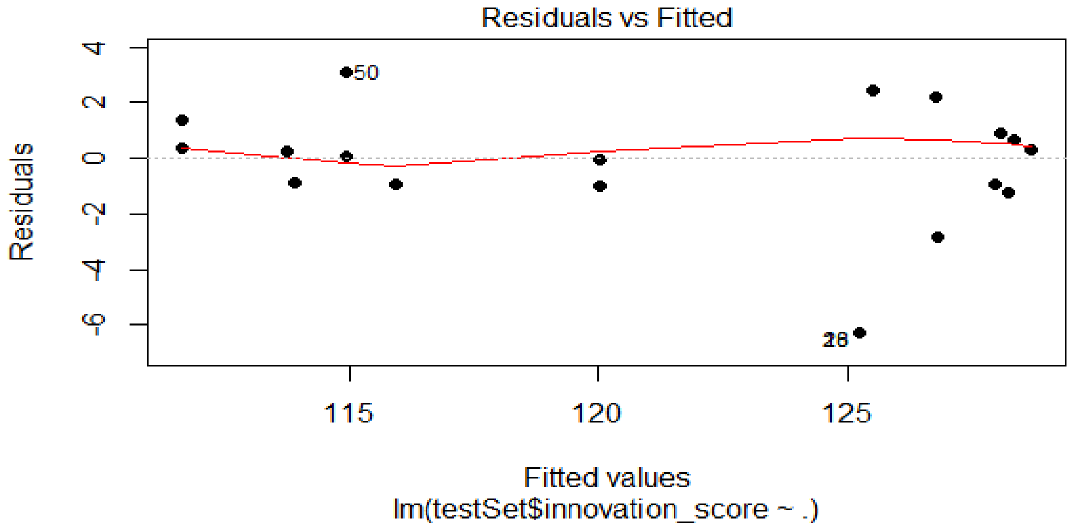


Figure 11– TestSet Regression Outcome

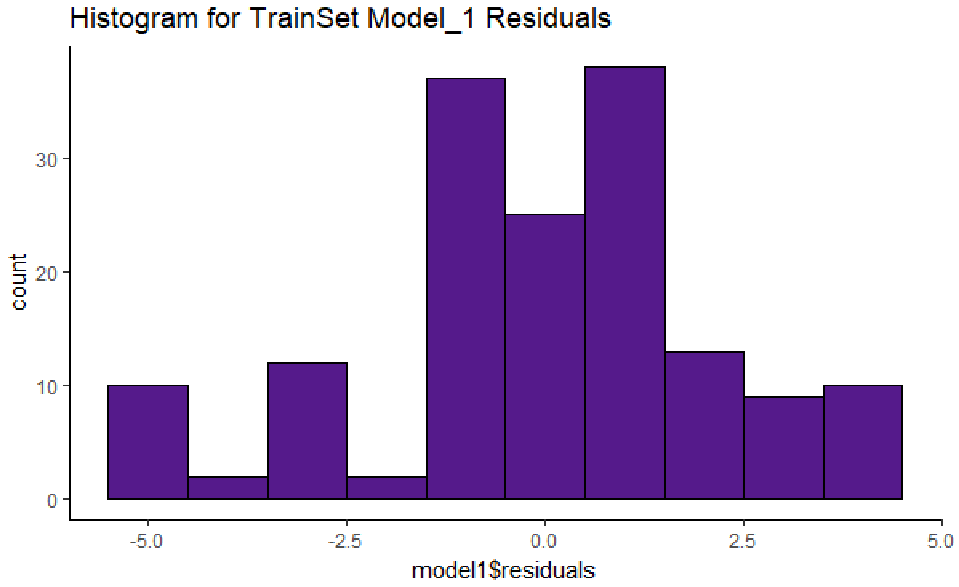


Figure 10 TrainSet Residual Distribution

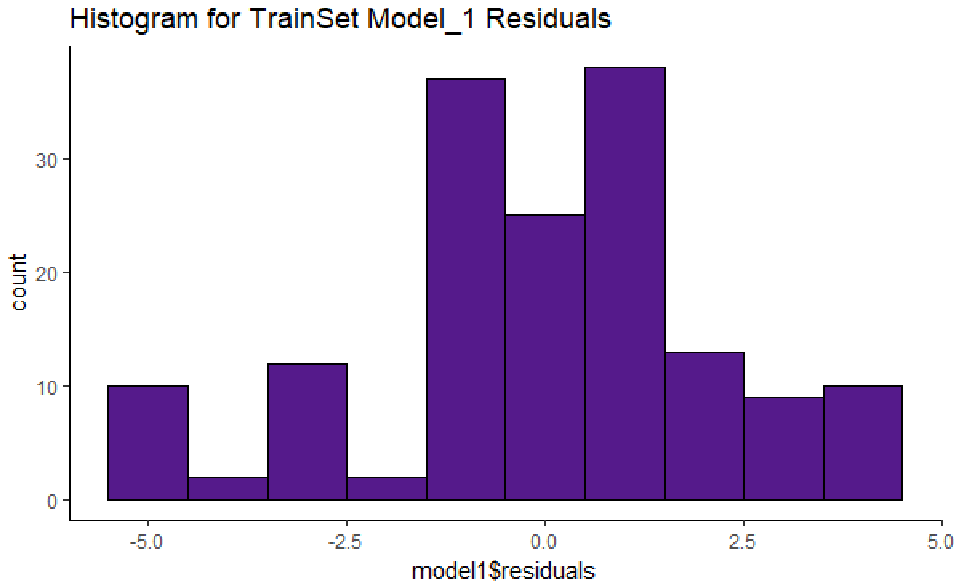


Figure 11 TestSet Residual Distribution

UNIVERSITIES DATA SET

KPIs->innovation_score, patent_application, scientists_and_engineers, public_rd, hgh_citations, value_added, education_expenditure
Since the data are loaded in R, transforming the variables is the next goal. It is necessary to create data frames, to execute operations on the variables. The dataset is called universities, as a first step is to omit the NA values from the dataset, also it is important to remove the columns that are not necessary e.g. uni_fact_id. The next step is to investigate if normality applies to the dataset. Through an inspection of the histograms the variables appear to be normally distributed, therefore the transformation of data is necessary to create a predictive model.

We can see from the model output that both girth and height are significantly related to volume, and that the model fits our data well.
For the multiple linear regression, the predictor variables are: teaching, research, citations and industry_income. The response variable is going to be international outlook.
The multi linear regression showed that the model accounted for 6% of the variance and all the investigated variables are slightly significant for the explanatory variable.

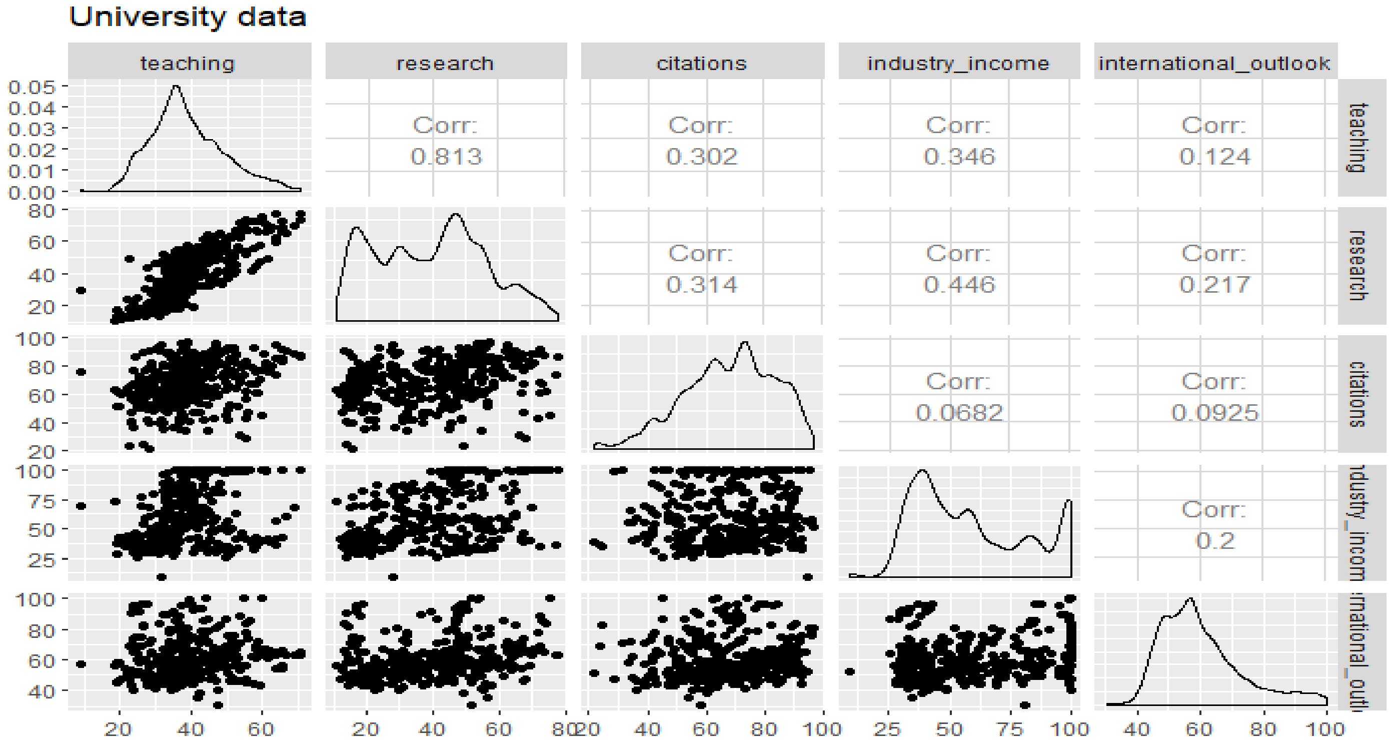


Figure 12 Variable Correlation

The matrix plot above visualizes the relationship among all variables in one single image. Fitting a linear regression on this dataset and see how well it models the observed data. Predictors will be added to give each of them a separate slope coefficient.

For our multiple linear regression example, we want to solve the following equation:

international outlook =B0+B1∗ teaching +B2∗ research +B3∗ citations + B4* industry_income

The model will estimate the value of the intercept (B0) and each predictor's slope (B1) for teaching, (B2) for research, (B3) citations, (B4) industry_income. The intercept is the expected international outlook for the value across all predictors. The value for each slope estimate is the increase in international outlook related with a one-unit increase in each predictor value, holding the others constant.

For better prediction purposes, the utilization of two subsets is necessary. The first is called trainSize with 75% of the observations and the other is called testSet with the remaining 25% of the observations.

The result of the models: Estimated international outlook is 52,8 in comparison to the training model which is 50,7 in the test model.

For improving one percentage point of research there will be an increase for international outlook by 2% in the training model, same with the test model. If there is an improvement in teaching by one point, it will lead for the training model to an extra 1,9 % decline for international outlook and 2 % decline for the test model. Note that R-squared number 6% for the training model 7% for the test model. Recalling the results from the two models, both had a significant p-value (close to zero). However, the most important fact is that F-statistic increased from 656 in the training model to 223 in the test model explaining the fact that there are more outliers in the former Results are available in Appendix C). Another important fact is that the p-value is less than 0,05(p-value: < 2.2e-16) for the training set and p-value: 8.173e-15 for the test set meaning that there is significant relationship between the variables in the dataset.

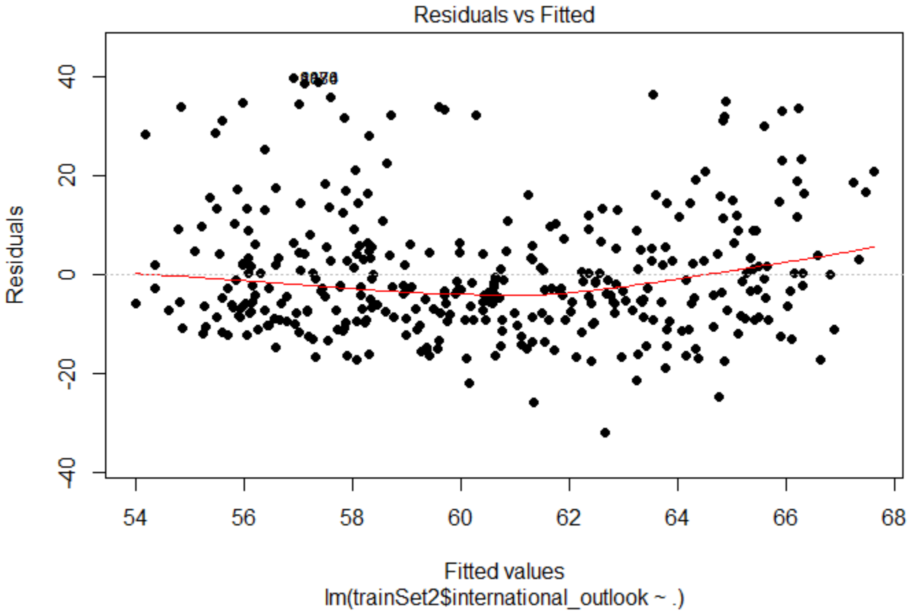


Figure 13 TrainSet2 Regression Outcome

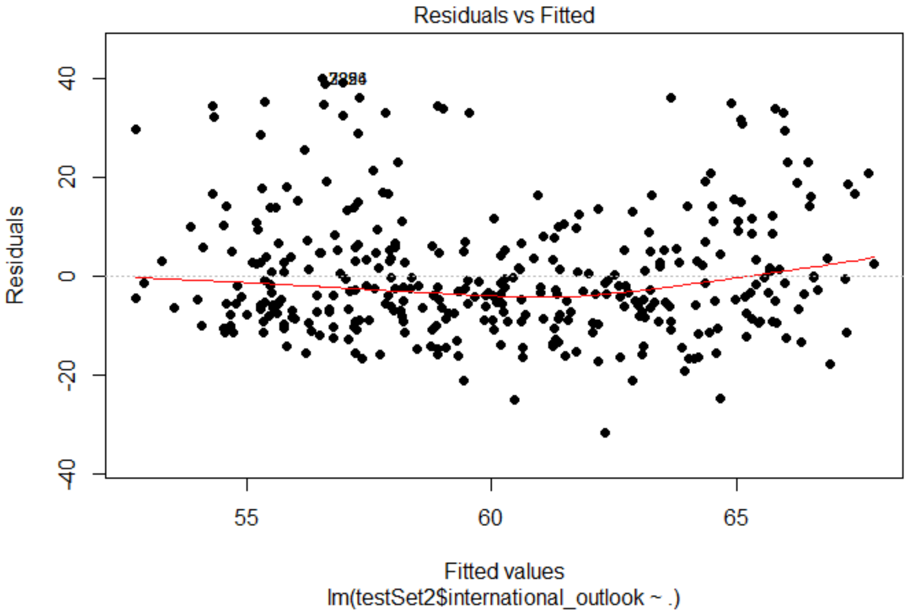


Figure 14 TestSet2 Regression Outcome

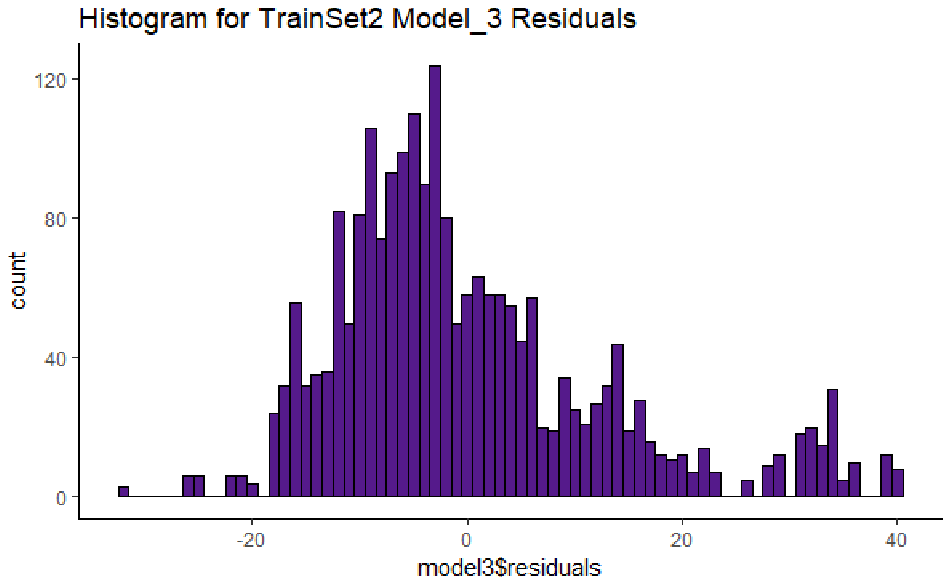


Figure 15 TrainSet2 Residuals Distribution

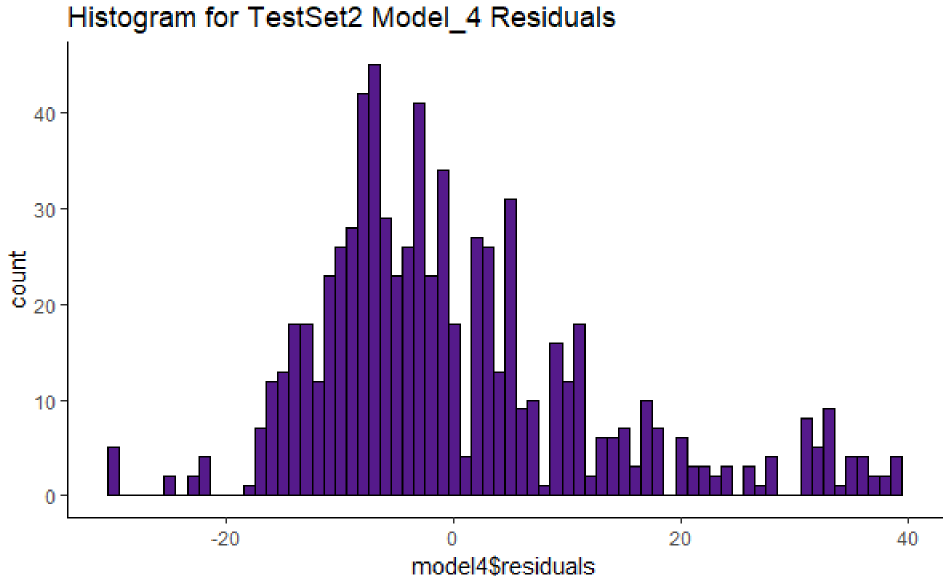
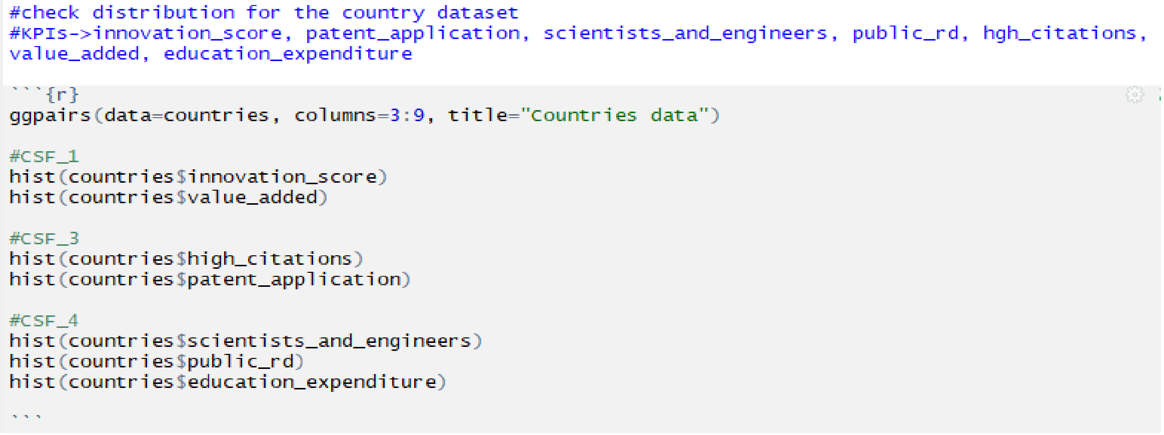


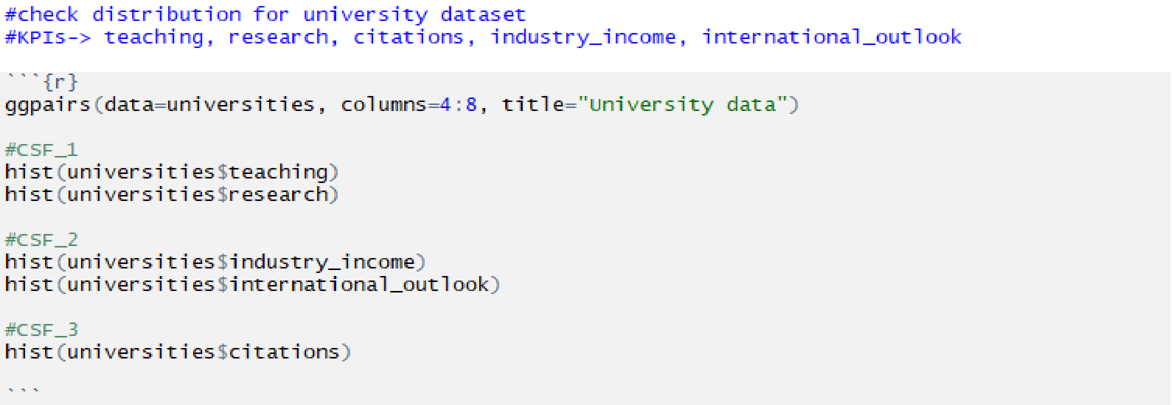
Figure 16 TestSet2 Residuals Distribution

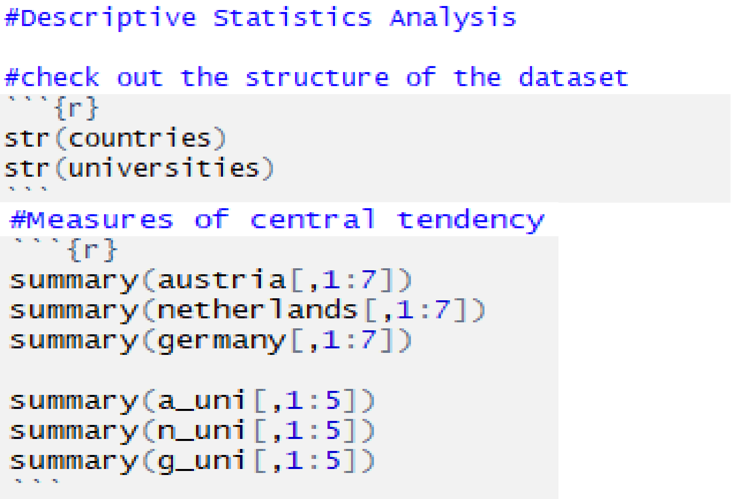
R code used for this section

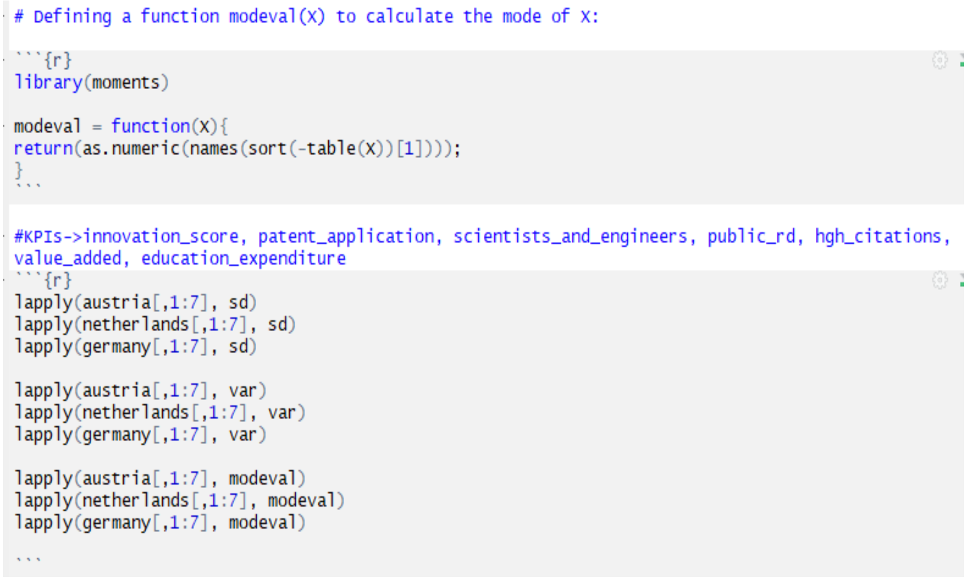


~~~
#check distribution for the country dataset
#KPIs->innovation_score, patent_application, scientists_and_engineers, public_rd, hgh_citations,
value_added, education_expenditure

```{r}
ggpairs(data=countries, columns=3:9, title="Countries data")

#CSF_1
hist(countries$innovation_score)
hist(countries$value_added)

#CSF_3
hist(countries$high_citations)
hist(countries$patent_application)

#CSF_4
hist(countries$scientists_and_engineers)
hist(countries$public_rd)
hist(countries$education_expenditure)
```
~~~

Figure 17 Correlation and Density

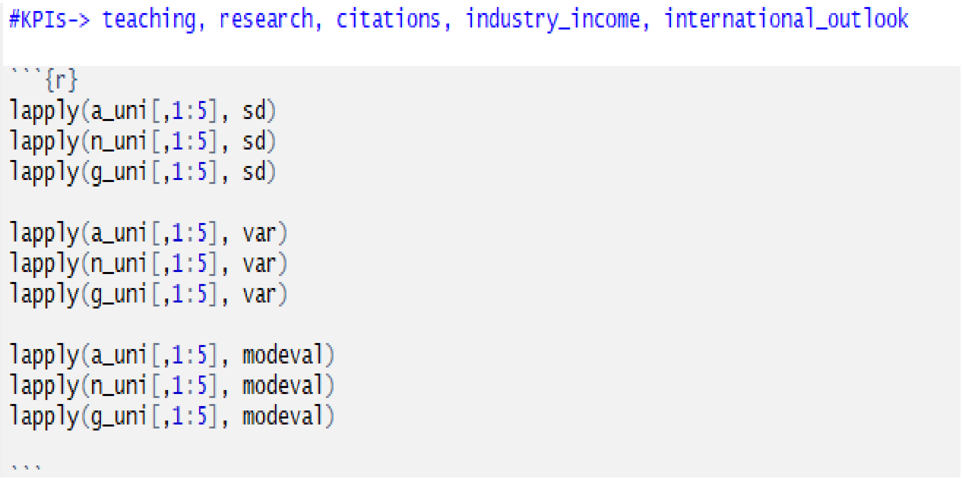


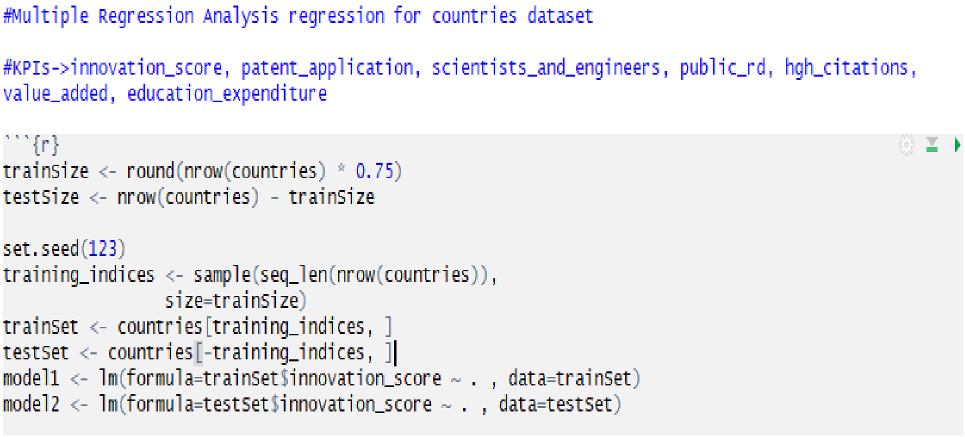
~~~
#check distribution for university dataset
#KPIs-> teaching, research, citations, industry_income, international_outlook

```{r}
ggpairs(data=universities, columns=4:8, title="University data")

#CSF_1
hist(universities$teaching)
hist(universities$research)

#CSF_2
hist(universities$industry_income)
hist(universities$international_outlook)

#CSF_3
hist(universities$citations)
~~~
```

Figure 18 Correlation and Density 2



```
#Descriptive Statistics Analysis

#check out the structure of the dataset
```{r}
str(countries)
str(universities)
```
#Measures of central tendency
```{r}
summary(austria[,1:7])
summary(netherlands[,1:7])
summary(germany[,1:7])

summary(a_uni[,1:5])
summary(n_uni[,1:5])
summary(g_uni[,1:5])
```
```

Figure 19 Structure and Summary



```
# Defining a function modeval(X) to calculate the mode of X:

```{r}
library(moments)

modeval = function(x){
return(as.numeric(names(sort(-table(x))[1])));
}
```

#KPIs->innovation_score, patent_application, scientists_and_engineers, public_rd, hgh_citations,
value_added, education_expenditure
```{r}
lapply(austria[,1:7], sd)
lapply(netherlands[,1:7], sd)
lapply(germany[,1:7], sd)

lapply(austria[,1:7], var)
lapply(netherlands[,1:7], var)
lapply(germany[,1:7], var)

lapply(austria[,1:7], modeval)
lapply(netherlands[,1:7], modeval)
lapply(germany[,1:7], modeval)

```
```

Figure 20 Standard Deviation, Variance and Mode
```

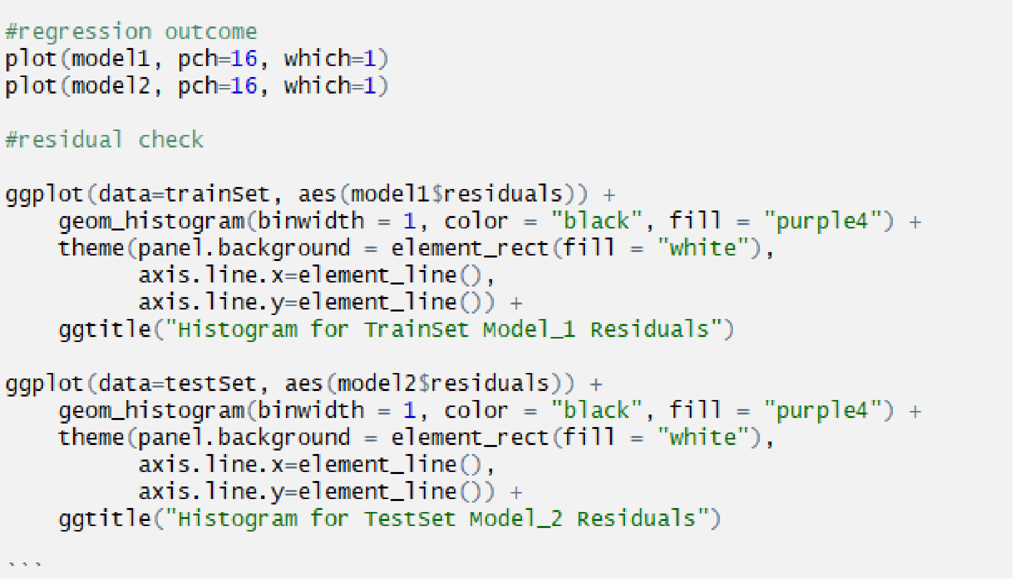


~~~
#KPIs-> teaching, research, citations, industry_income, international_outlook

```{r}
lapply(a_uni[,1:5], sd)
lapply(n_uni[,1:5], sd)
lapply(g_uni[,1:5], sd)

lapply(a_uni[,1:5], var)
lapply(n_uni[,1:5], var)
lapply(g_uni[,1:5], var)

lapply(a_uni[,1:5], modeval)
lapply(n_uni[,1:5], modeval)
lapply(g_uni[,1:5], modeval)

~~~
```

Figure 21 Standard Deviation, Variance and Mode 2



```
#Multiple Regression Analysis regression for countries dataset

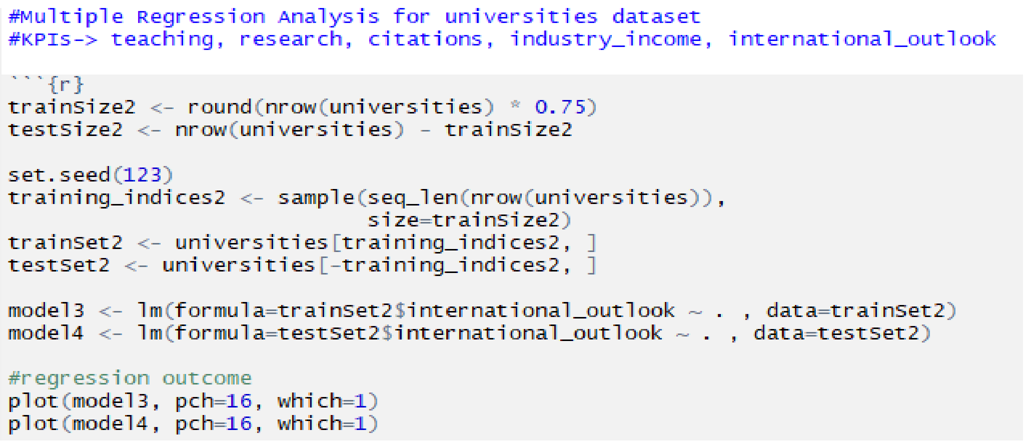
#KPIs->innovation_score, patent_application, scientists_and_engineers, public_rd, hgh_citations,
value_added, education_expenditure

```{r}
trainSize <- round(nrow(countries) * 0.75)
testSize <- nrow(countries) - trainSize

set.seed(123)
training_indices <- sample(seq_len(nrow(countries)),
                   size=trainSize)
trainSet <- countries[training_indices, ]
testSet <- countries[-training_indices, ]
model1 <- lm(formula=trainSet$innovation_score ~ . , data=trainSet)
model2 <- lm(formula=testSet$innovation_score ~ . , data=testSet)
```

Figure 22 Regression Analysis 1

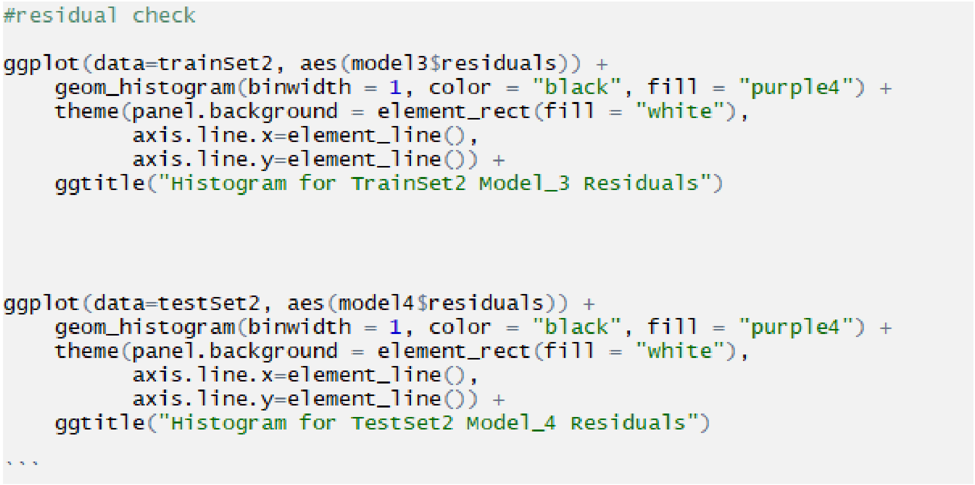


```
#regression outcome
plot(model1, pch=16, which=1)
plot(model2, pch=16, which=1)

#residual check

ggplot(data=trainSet, aes(model1$residuals)) +
    geom_histogram(binwidth = 1, color = "black", fill = "purple4") +
    theme(panel.background = element_rect(fill = "white"),
          axis.line.x=element_line(),
          axis.line.y=element_line()) +
    ggtitle("Histogram for TrainSet Model_1 Residuals")

ggplot(data=testSet, aes(model2$residuals)) +
    geom_histogram(binwidth = 1, color = "black", fill = "purple4") +
    theme(panel.background = element_rect(fill = "white"),
          axis.line.x=element_line(),
          axis.line.y=element_line()) +
    ggtitle("Histogram for TestSet Model_2 Residuals")

```
```

Figure 23 Regression Analysis Plots 1
```



~~~
#Multiple Regression Analysis for universities dataset
#KPIs-> teaching, research, citations, industry_income, international_outlook

```{r}
trainSize2 <- round(nrow(universities) * 0.75)
testSize2 <- nrow(universities) - trainSize2

set.seed(123)
training_indices2 <- sample(seq_len(nrow(universities)),
                            size=trainSize2)
trainSet2 <- universities[training_indices2, ]
testSet2 <- universities[-training_indices2, ]

model3 <- lm(formula=trainSet2$international_outlook ~ . , data=trainSet2)
model4 <- lm(formula=testSet2$international_outlook ~ . , data=testSet2)

#regression outcome
plot(model3, pch=16, which=1)
plot(model4, pch=16, which=1)
~~~

Figure 24 Regression Analysis 2

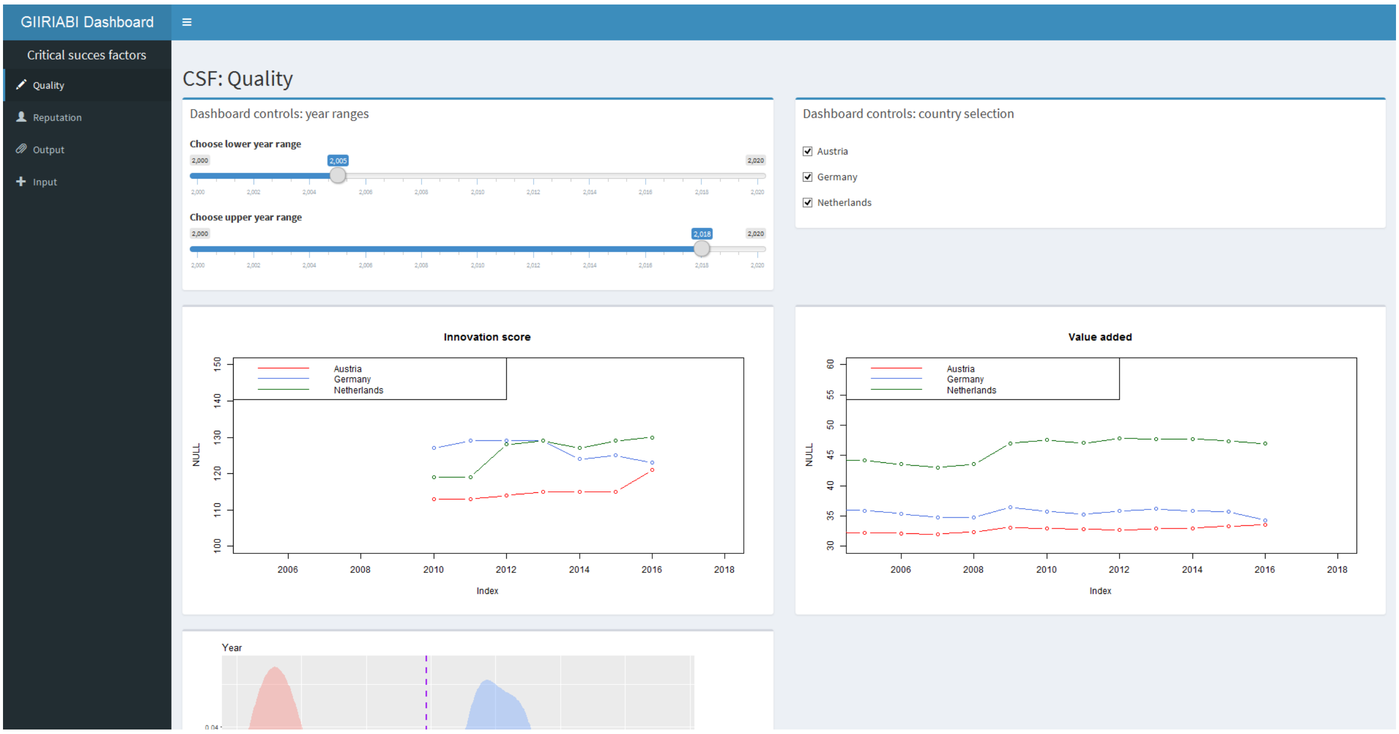


```
#residual check

ggplot(data=trainSet2, aes(model3$residuals)) +
    geom_histogram(binwidth = 1, color = "black", fill = "purple4") +
    theme(panel.background = element_rect(fill = "white"),
          axis.line.x=element_line(),
          axis.line.y=element_line()) +
    ggtitle("Histogram for TrainSet2 Model_3 Residuals")



ggplot(data=testSet2, aes(model4$residuals)) +
    geom_histogram(binwidth = 1, color = "black", fill = "purple4") +
    theme(panel.background = element_rect(fill = "white"),
          axis.line.x=element_line(),
          axis.line.y=element_line()) +
    ggtitle("Histogram for TestSet2 Model_4 Residuals")
```
```

Figure 25 Regression Analysis Plots 2
```

# Dashboard functionalities

## Dashboard overview

The dashboard consists of three main sections: the header, the side panel and the body. The header contains the text ‘GIIRIABI Dashboard’ in the top left corner and contains a button to hide or show the side panel. The side panel has five buttons to click on, four corresponding to a CSF and the fifth showing an overview of all CSFs for each country. The body section of the dashboard houses plots, graphs and dashboard controls to manipulate the plots and graphs. The body changes depending on which critical success factor is being inspected which is controlled by clicking on the buttons in the side panel. The body has a different layout for each critical success factor. Figure 26 and Figure 27 show how the dashboard looks and how the body can change depending on which critical success factor is being inspected.



Choose lower year range: 2,005 (2,000 – 2,020)

Choose upper year range: 2,018 (2,000 – 2,020)

☑ Austria

☑ Germany

☑ Netherlands

*Figure 26 – The “Quality” page of the dashboard*

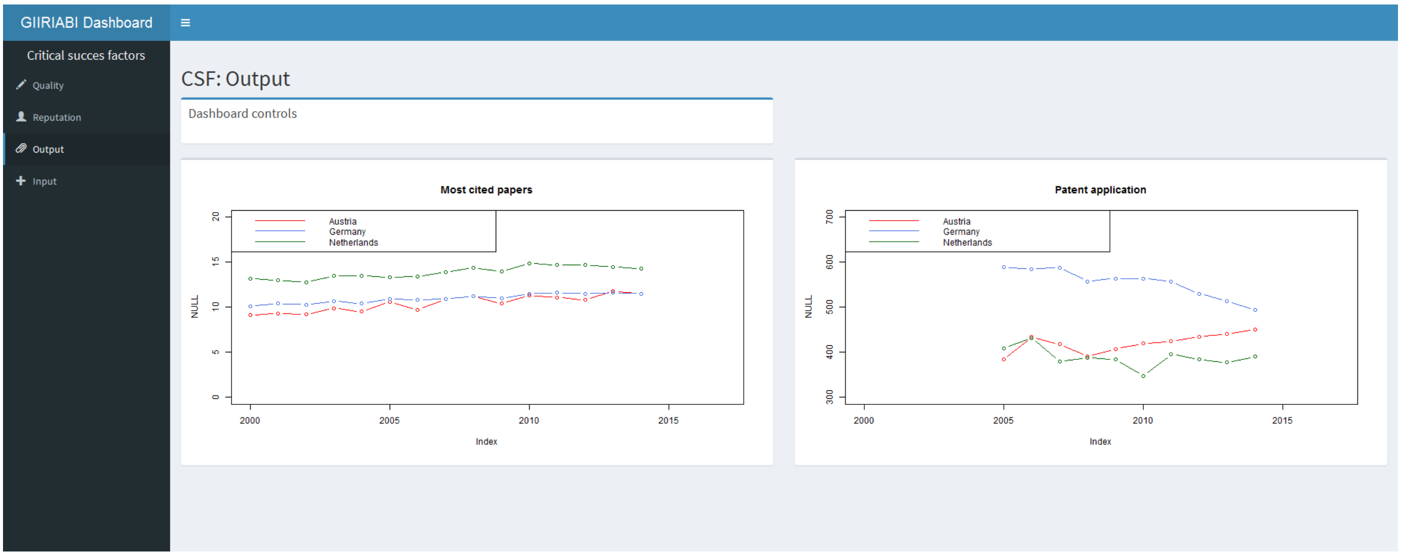
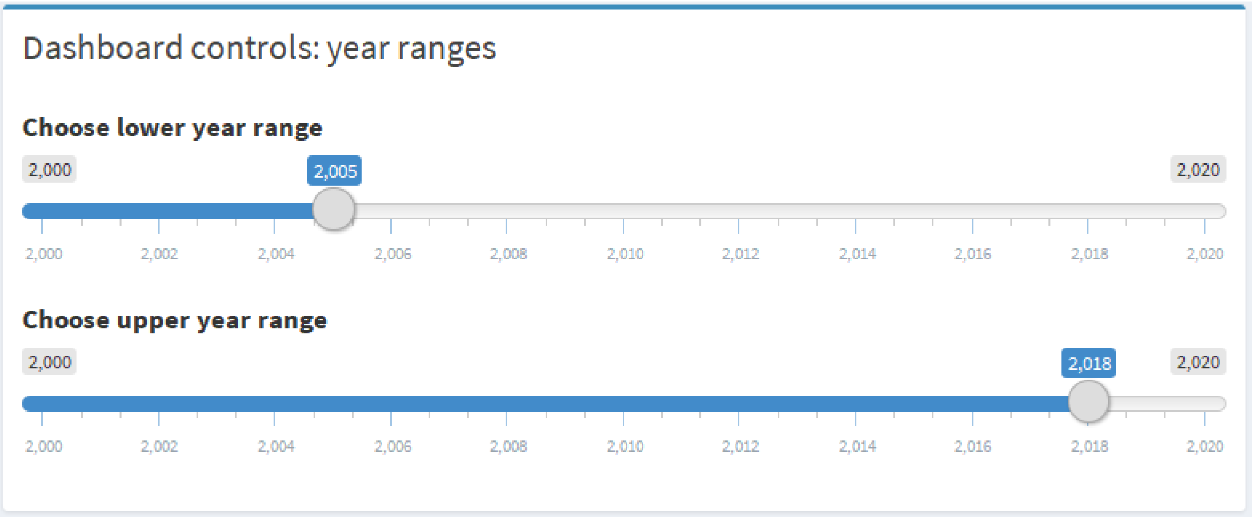


*Figure 27 – The “Output” page of the dashboard*

## Interaction

Each tab has some controls to interact with the data. The “year ranges” control allows the user to select from which years the data is shown. The standard display shows the data from all available years. The “year ranges” control can be found in Figure 28. An additional interaction is the selection of countries. By clicking on a country within the legend of a graph, that country can be deselected. The third major function is the ability to zoom in on data. Both these two functionalities are shown in Figure 29.



Dashboard controls: year ranges

Choose lower year range: 2,005

Choose upper year range: 2,018

*Figure 28 – The “year ranges” control*

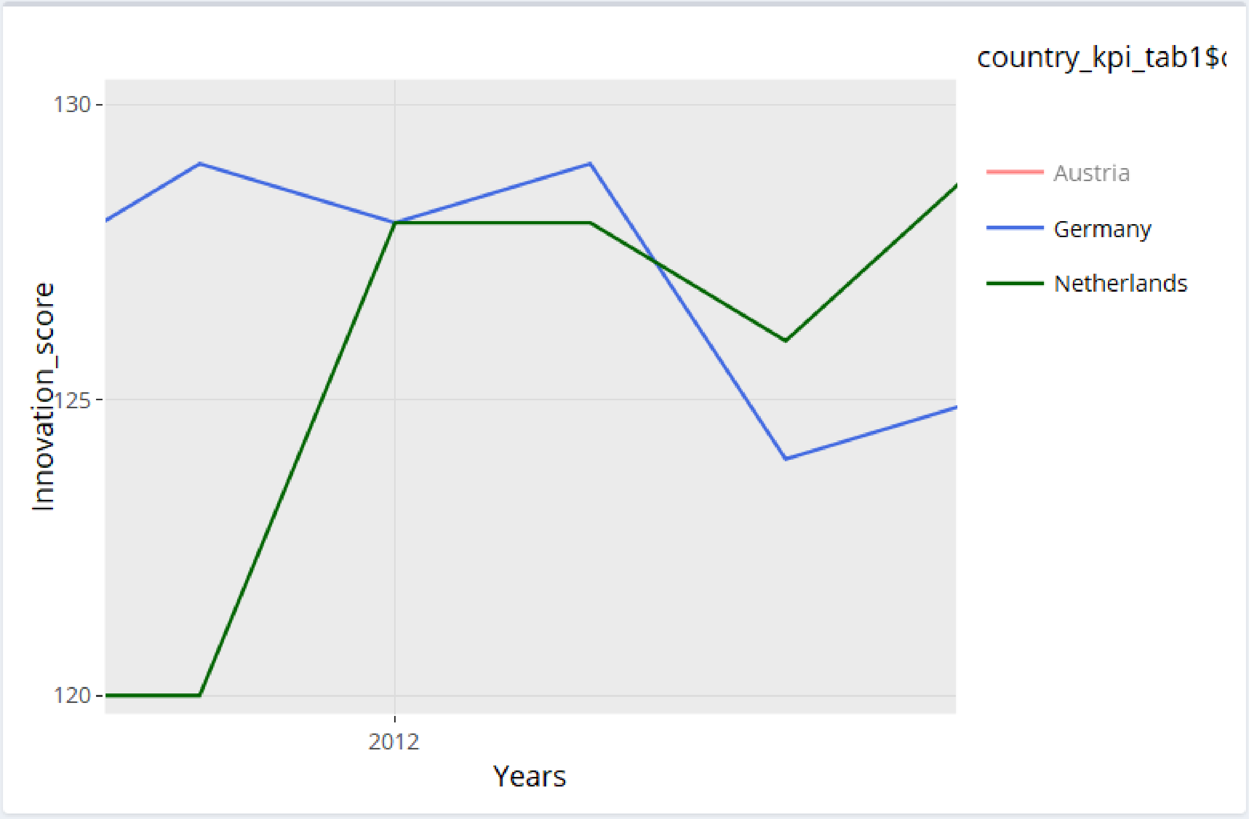


*Figure 29 – The innovation graph, where Austria has been deselected and the data has been zoomed in on*

## Performance thresholds

On the overview page, one can quickly see how each country has performed on each KPI for each CSF in the most recent year. In addition, one can see an interpretation of that performance, i.e. whether the country has performed well or not.

The following performance threshold categories are defined:

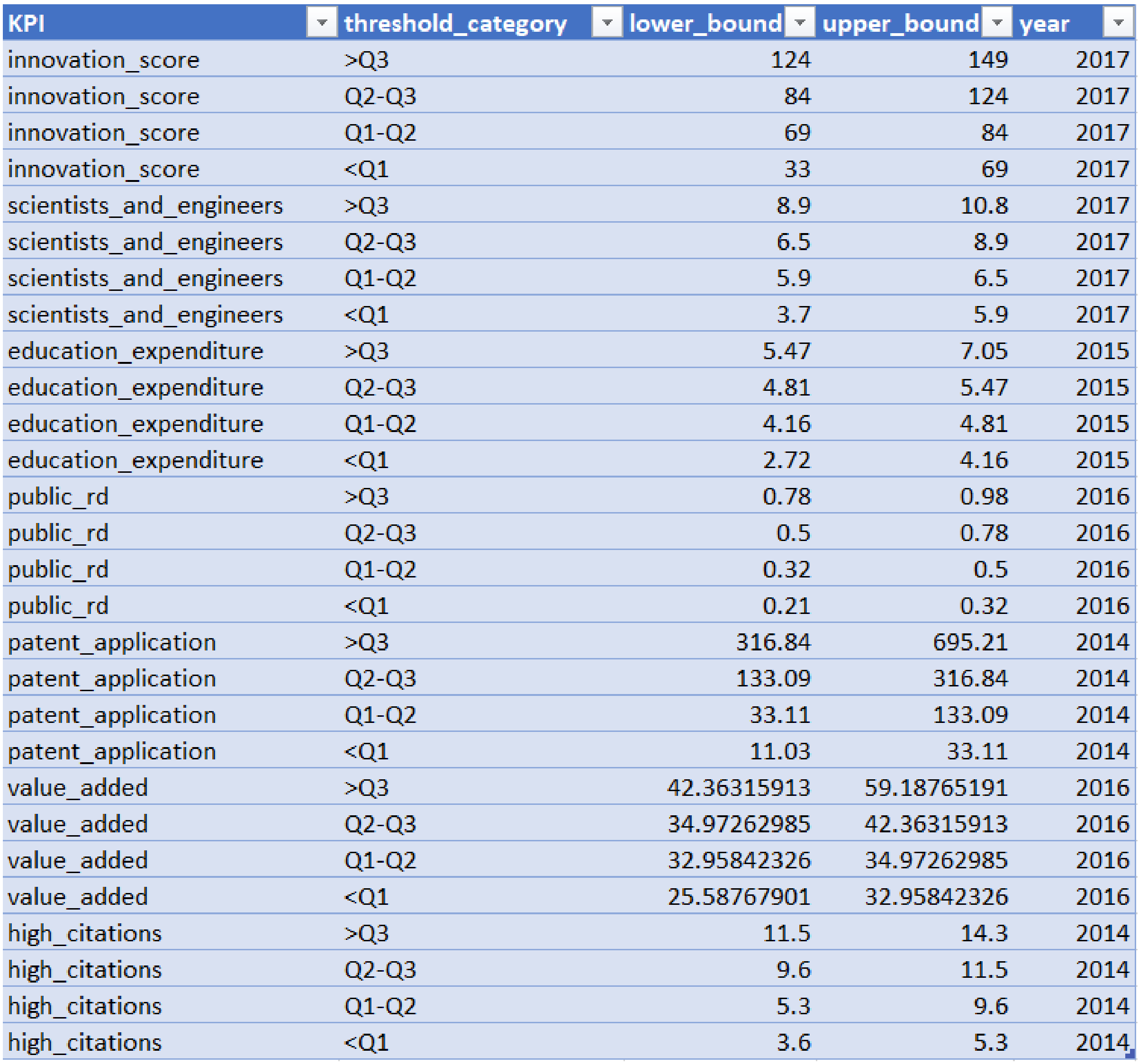
- Platinum: countries that perform relatively the best in terms of the KPI and can be considered an example to be followed by other countries.
- Gold: countries that perform well in terms of the KPI, but do not move horizons.
- Silver: countries that perform bad in terms of the KPI and need to improve their innovation efforts.
- Bronze: countries that perform dramatically bad in terms of the KPI and need to make structural changes to their innovation policy.

Ideally, a country will want to be in the Platinum category for each KPI, but the Gold category is good as well. If a country is in the Silver or Bronze category for a KPI then these countries are performing poorly. The calculations determining what category a country belongs to for a KPI are different for KPIs on the country dimension and for KPIs on the university dimensions.

The following KPIs are on the country dimension: innovation score, scientists and engineers, education expenditure, public R&D, patent application, value added, high citations. For these KPIs, data from the original datasets were re-extracted to get the most recent value for all available countries. Then, data for countries for which data was not available for each KPI were removed. The result of this operation was that 21 countries remained, which are in alphabetical order: Austria, Belgium, Bulgaria, Cyprus, Finland, France, Germany, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Poland, Romania, Slovakia, Slovenia, Spain, and Sweden. Using this data, the quartiles were calculated for each KPI. The performance thresholds are calcualted as follows:

- Platinum: the area between the highest value of the dataset and the third quartile.
- Gold: the area between the second and third quartile.
- Silver: the area between the first and second quartile.
- Bronze: the area between the lowest value of the dataset and the first quartile.

The quartiles for each KPI can be found in Figure 30. An Excel file with how these are calculated can be found on git.



| KPI | threshold_category | lower_bound | upper_bound | year |
|---|---|---|---|---|
| innovation_score | >Q3 | 124 | 149 | 2017 |
| innovation_score | Q2-Q3 | 84 | 124 | 2017 |
| innovation_score | Q1-Q2 | 69 | 84 | 2017 |
| innovation_score | <Q1 | 33 | 69 | 2017 |
| scientists_and_engineers | >Q3 | 8.9 | 10.8 | 2017 |
| scientists_and_engineers | Q2-Q3 | 6.5 | 8.9 | 2017 |
| scientists_and_engineers | Q1-Q2 | 5.9 | 6.5 | 2017 |
| scientists_and_engineers | <Q1 | 3.7 | 5.9 | 2017 |
| education_expenditure | >Q3 | 5.47 | 7.05 | 2015 |
| education_expenditure | Q2-Q3 | 4.81 | 5.47 | 2015 |
| education_expenditure | Q1-Q2 | 4.16 | 4.81 | 2015 |
| education_expenditure | <Q1 | 2.72 | 4.16 | 2015 |
| public_rd | >Q3 | 0.78 | 0.98 | 2016 |
| public_rd | Q2-Q3 | 0.5 | 0.78 | 2016 |
| public_rd | Q1-Q2 | 0.32 | 0.5 | 2016 |
| public_rd | <Q1 | 0.21 | 0.32 | 2016 |
| patent_application | >Q3 | 316.84 | 695.21 | 2014 |
| patent_application | Q2-Q3 | 133.09 | 316.84 | 2014 |
| patent_application | Q1-Q2 | 33.11 | 133.09 | 2014 |
| patent_application | <Q1 | 11.03 | 33.11 | 2014 |
| value_added | >Q3 | 42.36315913 | 59.18765191 | 2016 |
| value_added | Q2-Q3 | 34.97262985 | 42.36315913 | 2016 |
| value_added | Q1-Q2 | 32.95842326 | 34.97262985 | 2016 |
| value_added | <Q1 | 25.58767901 | 32.95842326 | 2016 |
| high_citations | >Q3 | 11.5 | 14.3 | 2014 |
| high_citations | Q2-Q3 | 9.6 | 11.5 | 2014 |
| high_citations | Q1-Q2 | 5.3 | 9.6 | 2014 |
| high_citations | <Q1 | 3.6 | 5.3 | 2014 |

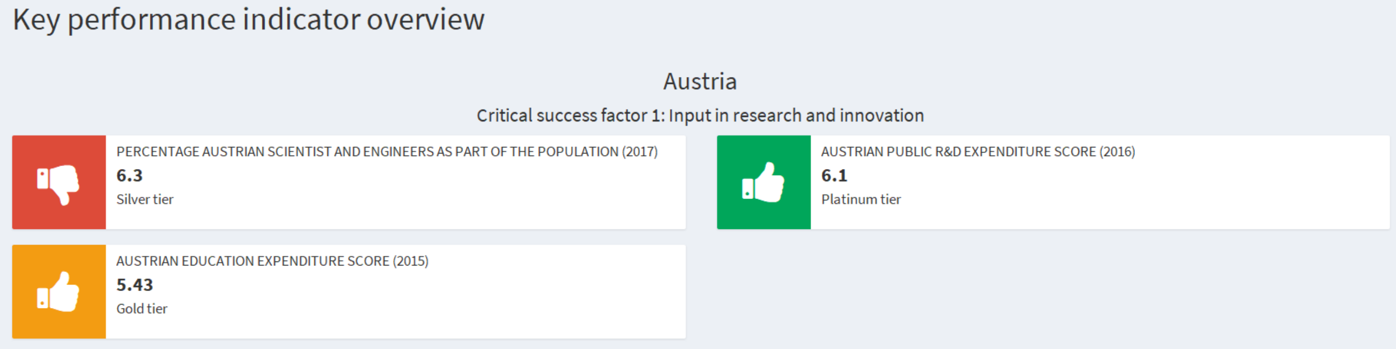
*Figure 30 – performance thresholds for country KPIs*

The following KPIs are on the university dimension: teaching, research, international outlook, industry, and citations. For these KPIs, data from the original datasets was explored using the source’s online tool, to find the value of universities for a KPI at a certain rank. For example, the teaching score of the university that is rank 100 in terms of teaching. These values are from the most recent year: 2018. The performance thresholds are calculated as follows:

- Platinum: the top 100 universities
- Gold: the top 200-101 universities
- Silver: the top 300-201 universities
- Bronze: universities below the top 300

As an additional visual aid, colours and thumbs are used to help the user see whether a value is good. A green colour with a thumbs up indicates the Platinum category, an orange colour with the thumbs up indicates the Gold category, a red colour with the thumbs down indicates the Silver category, and a black colour with the thumbs down indicates the Bronze category.

An excerpt of the key performance indicator overview page can be found in Figure 31. For each country an overview of all KPIs per CSF is shown, with categories and visual aids as discussed in this section.



Key performance indicator overview

Austria

Critical success factor 1: Input in research and innovation

👎 PERCENTAGE AUSTRIAN SCIENTIST AND ENGINEERS AS PART OF THE POPULATION (2017)
**6.3**
Silver tier

👍 AUSTRIAN PUBLIC R&D EXPENDITURE SCORE (2016)
**6.1**
Platinum tier

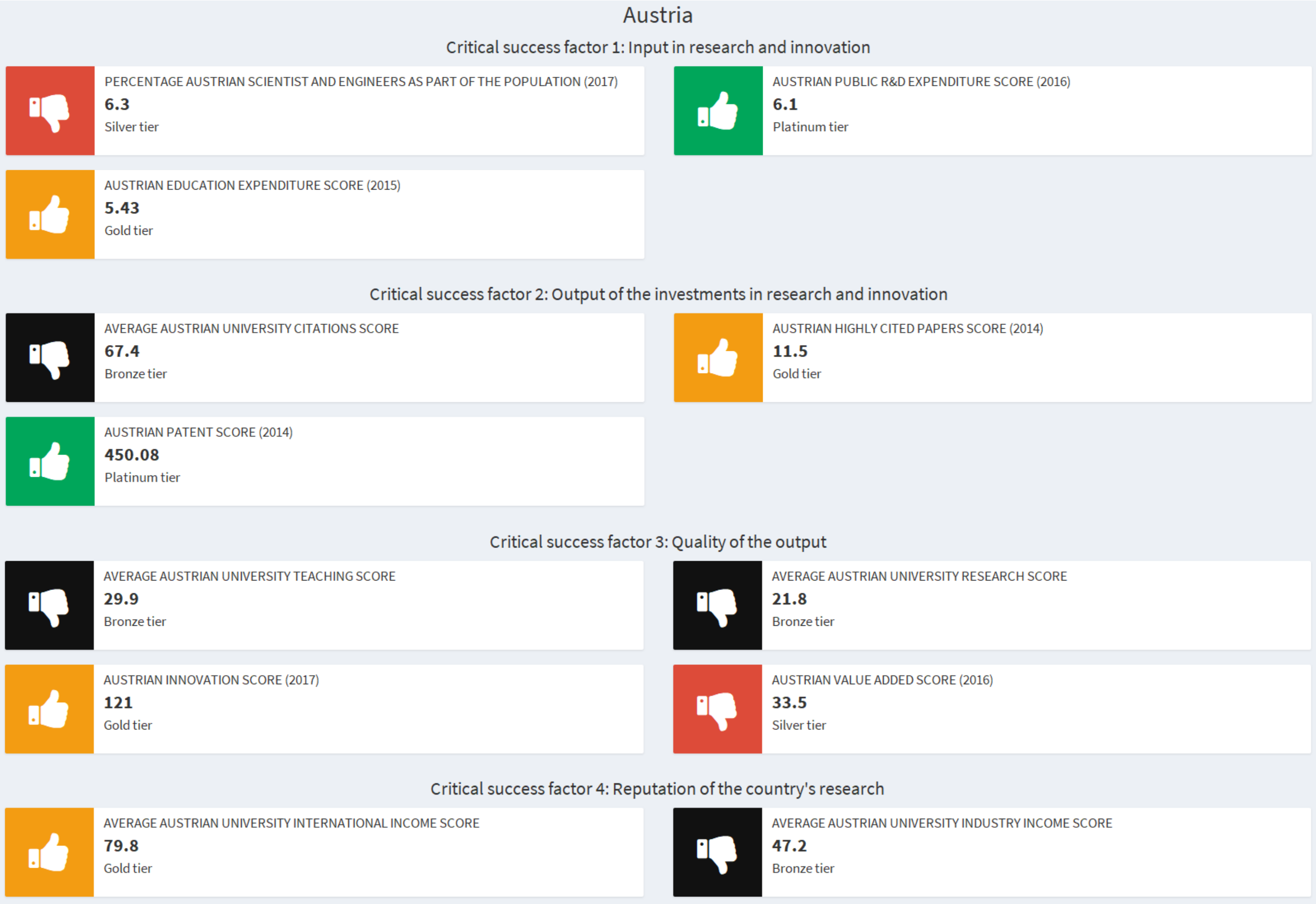
👍 AUSTRIAN EDUCATION EXPENDITURE SCORE (2015)
**5.43**
Gold tier

*Figure 31 – Key performance indicator overview page*

# Recommendations

From the dashboard Overview page several  conclusions can be drawn for each country.

Austria has achieved very mixed results as can be seen in Figure 32, ranging from achieving Platinum category in some places and Bronze categories in others. Austria has achieved Platinum category for public R&D expenditure, but given the poor scores all across the board Austria should assess whether or not these expenditures are being used effectively. Another KPI Austria has achieved Platinum category for is the number of patents, so Austria should continue stimulating patents the way there are currently doing. The most distressing KPIs are those relating to the KPIs on the university dimension, as four of the five KPIs have achieved a Bronze category. Given that Austria has achieved a Gold category on the education expenditure KPI, Austria should assess whether those expenses are being used effectively.



Austria

Critical success factor 1: Input in research and innovation

👎 PERCENTAGE AUSTRIAN SCIENTIST AND ENGINEERS AS PART OF THE POPULATION (2017)
**6.3**
Silver tier

👍 AUSTRIAN PUBLIC R&D EXPENDITURE SCORE (2016)
**6.1**
Platinum tier

👍 AUSTRIAN EDUCATION EXPENDITURE SCORE (2015)
**5.43**
Gold tier

Critical success factor 2: Output of the investments in research and innovation

👎 AVERAGE AUSTRIAN UNIVERSITY CITATIONS SCORE
**67.4**
Bronze tier

👍 AUSTRIAN HIGHLY CITED PAPERS SCORE (2014)
**11.5**
Gold tier

👍 AUSTRIAN PATENT SCORE (2014)
**450.08**
Platinum tier

Critical success factor 3: Quality of the output

👎 AVERAGE AUSTRIAN UNIVERSITY TEACHING SCORE
**29.9**
Bronze tier

👎 AVERAGE AUSTRIAN UNIVERSITY RESEARCH SCORE
**21.8**
Bronze tier

👍 AUSTRIAN INNOVATION SCORE (2017)
**121**
Gold tier

👎 AUSTRIAN VALUE ADDED SCORE (2016)
**33.5**
Silver tier

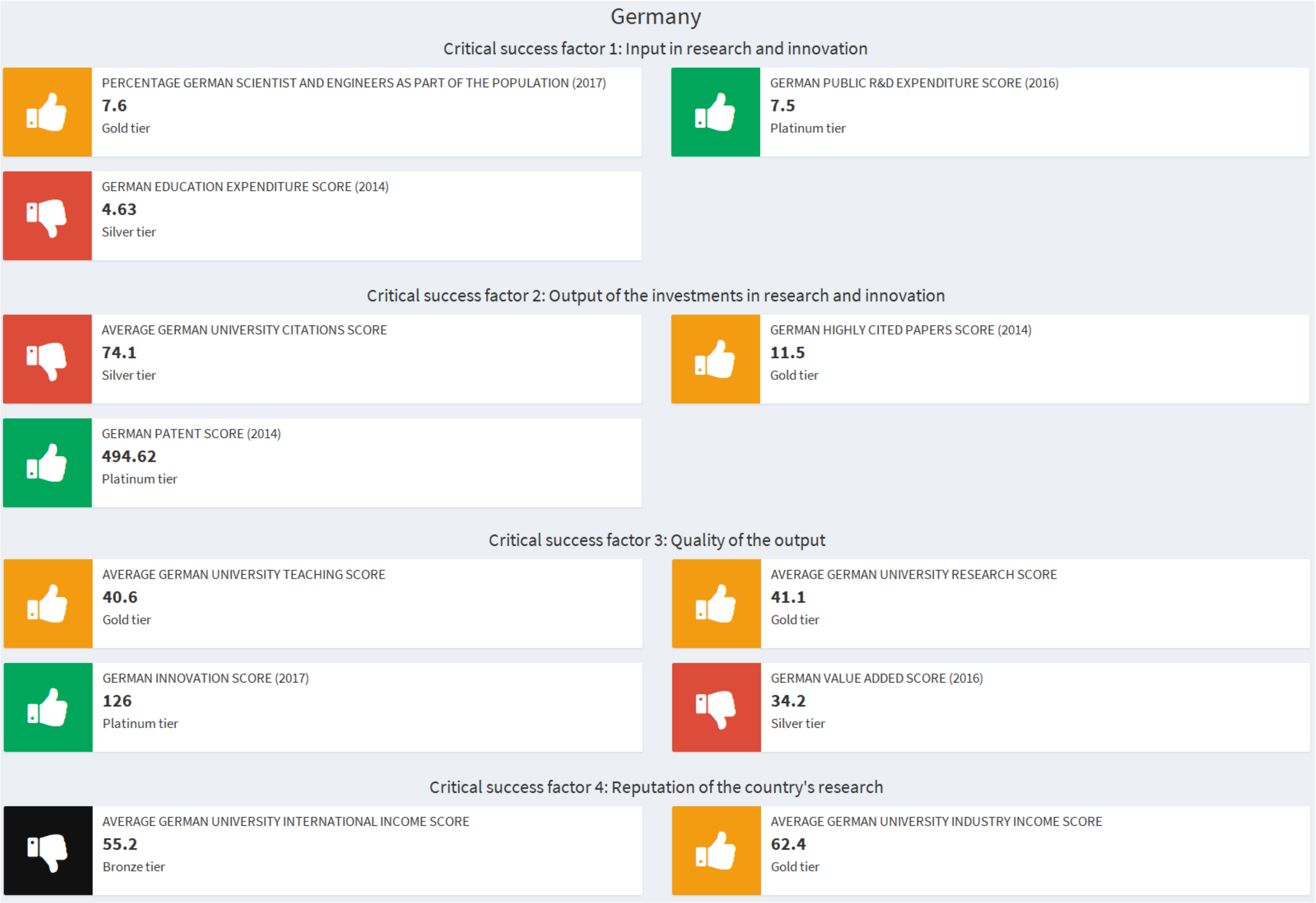
Critical success factor 4: Reputation of the country's research

👍 AVERAGE AUSTRIAN UNIVERSITY INTERNATIONAL INCOME SCORE
**79.8**
Gold tier

👎 AVERAGE AUSTRIAN UNIVERSITY INDUSTRY INCOME SCORE
**47.2**
Bronze tier

*Figure 32 – Austria performance thresholds*

Germany has performed decently as can be seen in Figure 33, with only a few categories on the lower half of the quartiles. The KPIs for which poor performance categories have been achieved (Bronze and Silver) are scattered amongst the different CSFs, making it difficult to formulate a precise policy advice for Germany. For starts, Germany should analyse the composition of the R&D expenditure, for which it has achieved a Platinum category, to see if enough of it is going to the KPIs for which Silver or Bronze categories have been achieved.



Germany

Critical success factor 1: Input in research and innovation

👍 PERCENTAGE GERMAN SCIENTIST AND ENGINEERS AS PART OF THE POPULATION (2017)
**7.6**
Gold tier

👍 GERMAN PUBLIC R&D EXPENDITURE SCORE (2016)
**7.5**
Platinum tier

👎 GERMAN EDUCATION EXPENDITURE SCORE (2014)
**4.63**
Silver tier

Critical success factor 2: Output of the investments in research and innovation

👎 AVERAGE GERMAN UNIVERSITY CITATIONS SCORE
**74.1**
Silver tier

👍 GERMAN HIGHLY CITED PAPERS SCORE (2014)
**11.5**
Gold tier

👍 GERMAN PATENT SCORE (2014)
**494.62**
Platinum tier

Critical success factor 3: Quality of the output

👍 AVERAGE GERMAN UNIVERSITY TEACHING SCORE
**40.6**
Gold tier

👍 AVERAGE GERMAN UNIVERSITY RESEARCH SCORE
**41.1**
Gold tier

👍 GERMAN INNOVATION SCORE (2017)
**126**
Platinum tier

👎 GERMAN VALUE ADDED SCORE (2016)
**34.2**
Silver tier

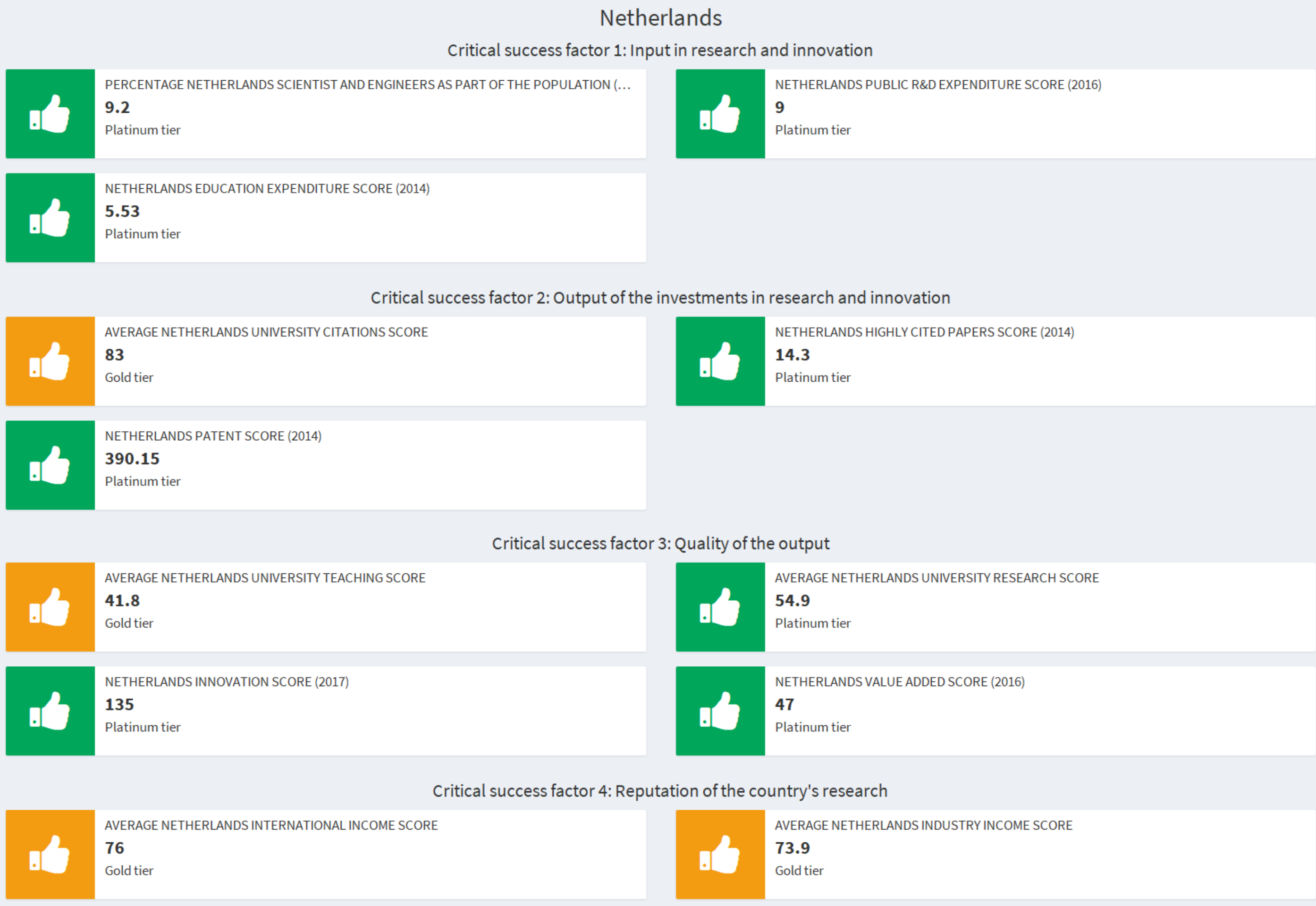
Critical success factor 4: Reputation of the country's research

👎 AVERAGE GERMAN UNIVERSITY INTERNATIONAL INCOME SCORE
**55.2**
Bronze tier

👍 AVERAGE GERMAN UNIVERSITY INDUSTRY INCOME SCORE
**62.4**
Gold tier

*Figure 33 – Germany performance thresholds*

The Netherlands is performing well all across the board as can be seen in Figure 34. The Netherlands is for most KPIs among the top 25% of all included countries and among the top 50% for all KPIs. A weak point of the Netherlands are the universities, as four of the five KPIs of the university dimension achieve a Gold category, whereas all KPIs of the country dimension achieve a Platinum category. The Netherlands should thus adopt two strategies: maintain leadership in the KPIs with a Platinum category and increase investments in the universities.



Netherlands

Critical success factor 1: Input in research and innovation

PERCENTAGE NETHERLANDS SCIENTIST AND ENGINEERS AS PART OF THE POPULATION (…
**9.2**
Platinum tier

NETHERLANDS PUBLIC R&D EXPENDITURE SCORE (2016)
**9**
Platinum tier

NETHERLANDS EDUCATION EXPENDITURE SCORE (2014)
**5.53**
Platinum tier

Critical success factor 2: Output of the investments in research and innovation

AVERAGE NETHERLANDS UNIVERSITY CITATIONS SCORE
**83**
Gold tier

NETHERLANDS HIGHLY CITED PAPERS SCORE (2014)
**14.3**
Platinum tier

NETHERLANDS PATENT SCORE (2014)
**390.15**
Platinum tier

Critical success factor 3: Quality of the output

AVERAGE NETHERLANDS UNIVERSITY TEACHING SCORE
**41.8**
Gold tier

AVERAGE NETHERLANDS UNIVERSITY RESEARCH SCORE
**54.9**
Platinum tier

NETHERLANDS INNOVATION SCORE (2017)
**135**
Platinum tier

NETHERLANDS VALUE ADDED SCORE (2016)
**47**
Platinum tier

Critical success factor 4: Reputation of the country's research

AVERAGE NETHERLANDS INTERNATIONAL INCOME SCORE
**76**
Gold tier

AVERAGE NETHERLANDS INDUSTRY INCOME SCORE
**73.9**
Gold tier

*Figure 34 – The Netherlands performance thresholds*

## Improvements

Given a successful pilot run, several improvements will be made to improve the quality of the business intelligence dashboard. First, improvements pertaining to the data used are discussed. Second, an improvement to comparisons is presented. Then, concrete plans for improving the performance thresholds are mentioned. Finally, an open issue on measuring success is presented.

An obvious improvement to the dashboard would be to include more countries. Indeed, this is one of the biggest plans, however there are interesting facets to this improvement to discuss. Most notably is that data collection will be a hurdle, as data is not available for all countries, not even for all countries from the European Union. As the number of countries increases, so does the quality of the performance thresholds (which are based on quartiles). This is important because performance thresholds are an indication of how a country is doing compared to other countries. Additionally, data can be collected for KPIs in different ways. For example, education expenditure as % of GDP can be used, but the absolute number can be used as well. All forms in which the data can be found should be supported by the dashboard, so that the user can view the data in the way in which they desire.

Currently, the user can compare three countries. Naturally, as the number of countries increase, so does the ability of the user to compare countries. But an additional feature is going to be implemented. More dimensions are going to be added on which data can be compared: continents and economic partnerships (e.g. EU). For example, a country from the EU may want to know how it compares to the EU average. This will be possible with the added data dimensions.

The performance threshold feature has some planned improvements for the future. First, the calculation for the threshold categories for KPIs on the university dimension might have to be changed, because they are currently decided on arbitrarily. The calculation for the KPIs on the country dimension is more solid, because quartiles form an understandable basis to calculate threshold categories from. But threshold categories such as the top 100 universities are not. Ideally, a partnership with the Times Higher Education (the source of the data for the KPIs on the university dimension) is formed that allows us to easily access the entire dataset, and thus calculate the quartiles for the universities as well. Second, something that was not available with the time resources in this phase of the project was a more precise calculation for the KPIs on the university dimension. Currently, the average of all universities of a country is used to determine to which threshold category a country belongs to. However, this is flawed as outliers can easily affect the average. Instead, the plan is to calculate for each university individually to which threshold category it belongs, and then show how much universities belong to each category. This way, the user can also easily see which universities need to be improved the most. Finally, there is a plan to allow for user-defined performance thresholds. The current performance thresholds categories are more long-term goals: a country will want to move to the Platinum category, but it is unlikely that it can be achieved in one year, especially for countries belonging to the Bronze category. Instead, users may want to define a goal to reach within a year, or within a number of year. For example, a country with an education expenditure of 1 may want to increase that to 1.5 within a year. This will not move the country out of the Bronze category, but it is an improvement still.

Finally, there exists an open issue on how to measure success. In this pilot run of the dashboard, success in research and innovation was measured by using four CSFs and several KPIs for each critical success factor. Practice and time will tell how exhaustive these CSFs and KPIs are, to what degree they may need to be adjusted, or what should be added or removed. The comparisons made in the Recommendations section indicate that the composition of the CSFs may need to be adjusted, or that more KPIs need to be added to each CSF, as for Austria and Germany the KPIs for each CSF often given conflicting measures of success.

# References

Mannila, H., & Smyth, P. Principles of Data Mining. By David Hand.

http://ec.europa.eu/invest-in-research/index_en.htm

https://ec.europa.eu/research/index.cfm?pg=newsalert&year=2017&na=na-030717

http://ec.europa.eu/growth/industry/innovation/facts-figures/scoreboards/

## Appendix A – Data sources

The data sources can be found in Table 1.

| CSF | KPI | Source |
|---|---|---|
| **Input in research and innovation** | Public research & development expenditure | https://rio.jrc.ec.europa.eu/en/stats/public-government-and-higher-education-rd-expenditure-gdp |
| | Human resources | https://rio.jrc.ec.europa.eu/en/stats/human-resources-science-and-technology-hrst-sub-groups |
| | Education expenditure | https://rio.jrc.ec.europa.eu/en/stats/expenditure-education-gdp-education-level |
| **Output of the investments in research and innovation** | Citations | https://www.timeshighereducation.com/world-university-rankings/2018/world-ranking#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats |
| | Highly cited papers | https://rio.jrc.ec.europa.eu/en/stats/highly-cited-publications |
| | Number of patent applications | http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=pat_ep_ntot&lang=en |
| **Quality of the output** | Quality of teaching | https://www.timeshighereducation.com/world-university-rankings/2018/world-ranking#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats |
| | Research | https://www.timeshighereducation.com/world-university-rankings/2018/world-ranking#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats |
| | Innovation score | http://ec.europa.eu/growth/industry/innovation/facts-figures/scoreboards/ |
| | Value added | https://rio.jrc.ec.europa.eu/en/stats/value-added-services-knowledge-intensity-total-value-added-services-and-total-value-added |
| **Reputation of the country's research** | International outlook | https://www.timeshighereducation.com/world-university-rankings/2018/world-ranking#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats |
| | Industry income | https://www.timeshighereducation.com/world-university-rankings/2018/world-ranking#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats |

*Table 1 – Data sources*

# Appendix B – Descriptive Statistic-Predictive Statistics Results and Plots

## Density Histograms for Countries dataset per Variable

KPIs->innovation_score, patent_application, scientists_and_engineers, public_rd, hgh_citations, value_added, education_expenditure

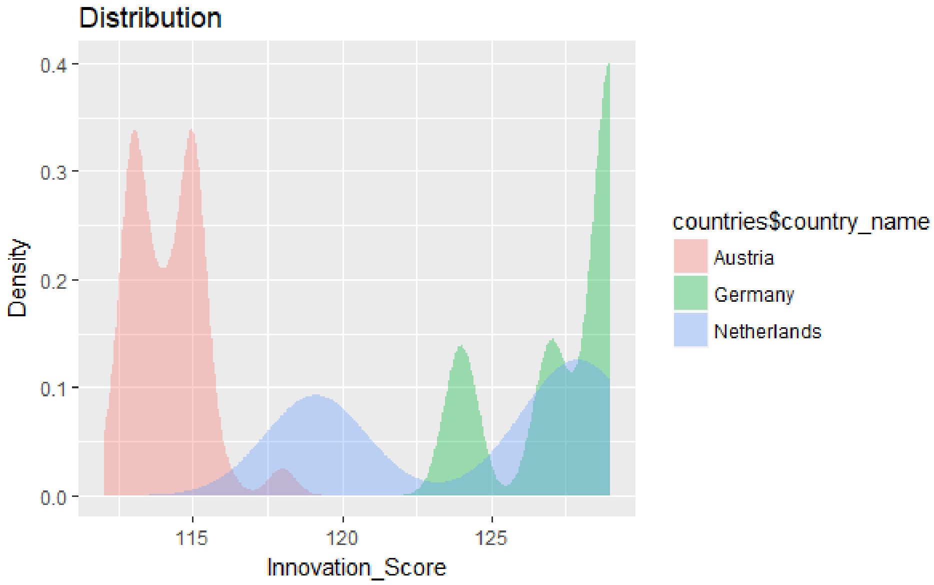


*Figure 35 Appendix– Density for Countries Innovation score*

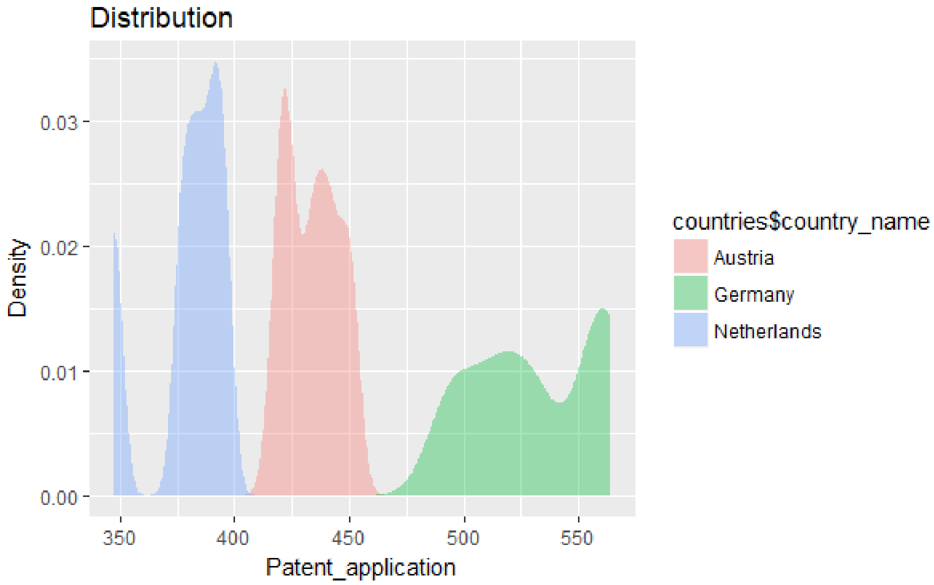


*Figure 36 Density for Countries patent application*

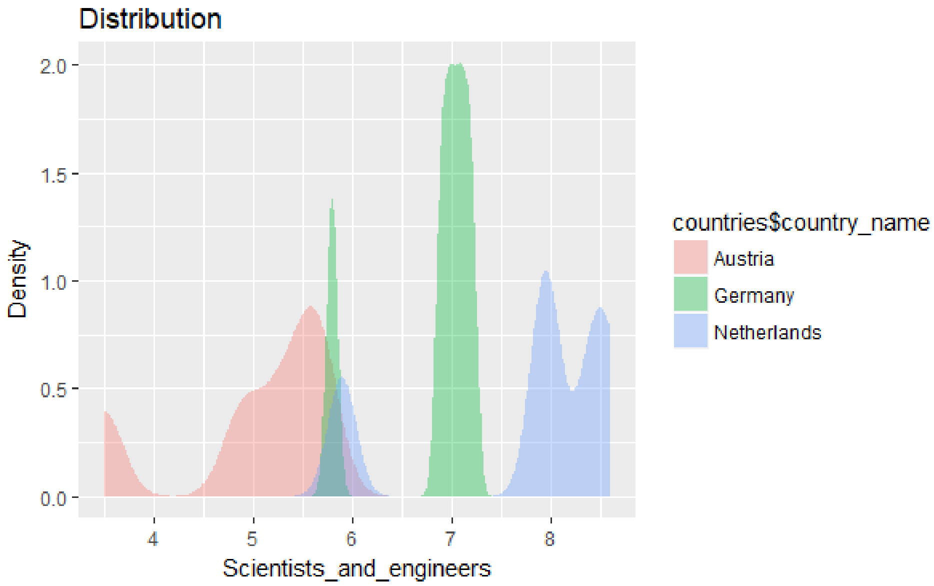


*Figure 37 Density for Countries scientists and engineers*

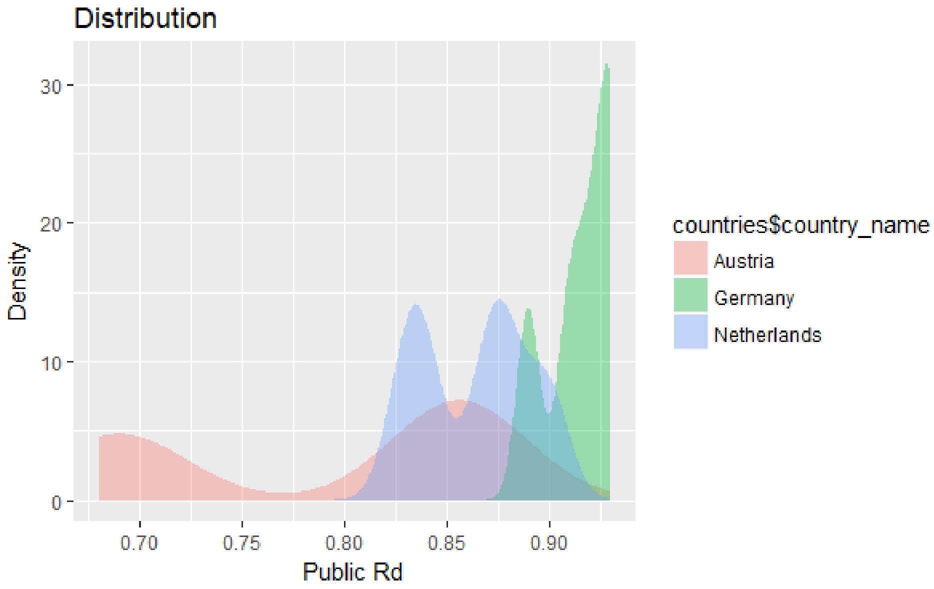


*Figure 38 Density for Countries public rd*

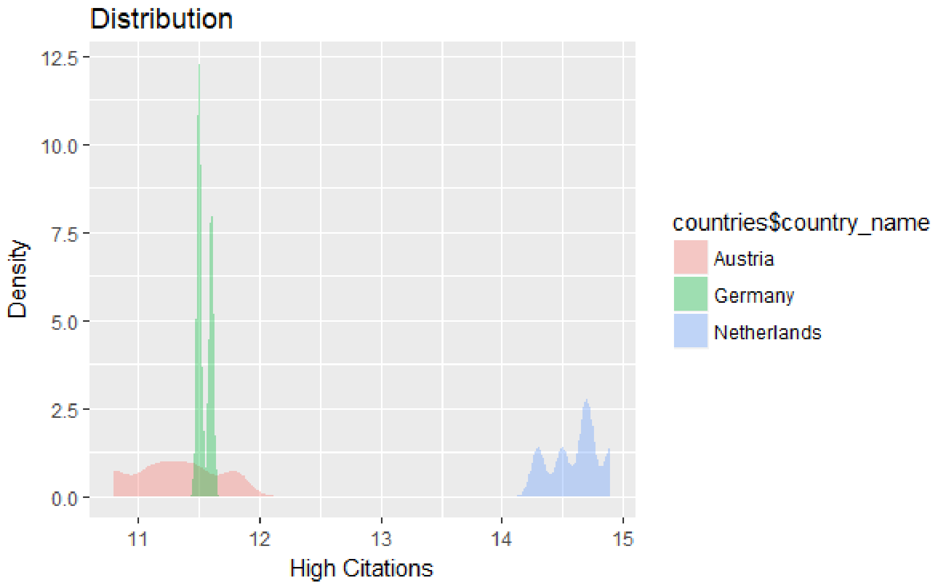


*Figure 39 Density for Countries high_citations*

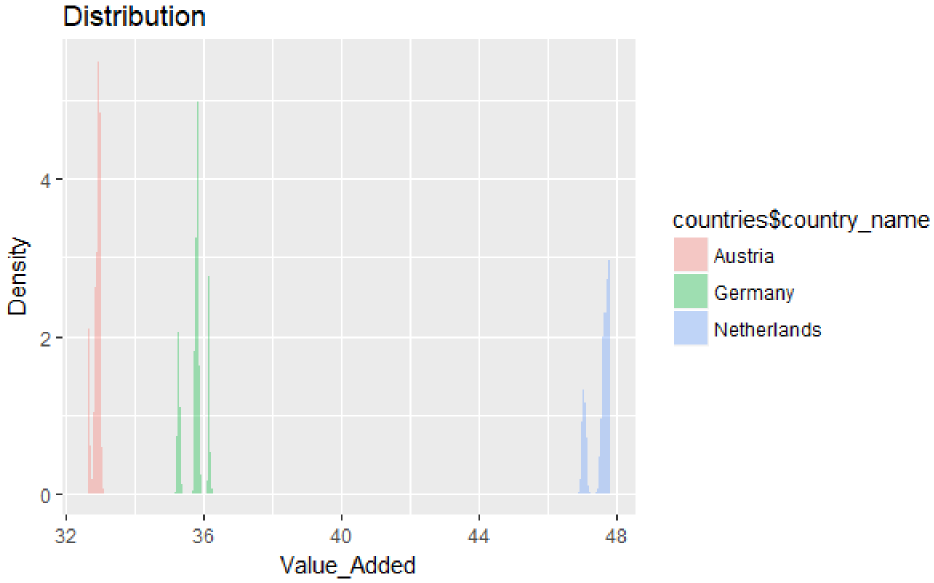


*Figure 40 Density for Countries value added*

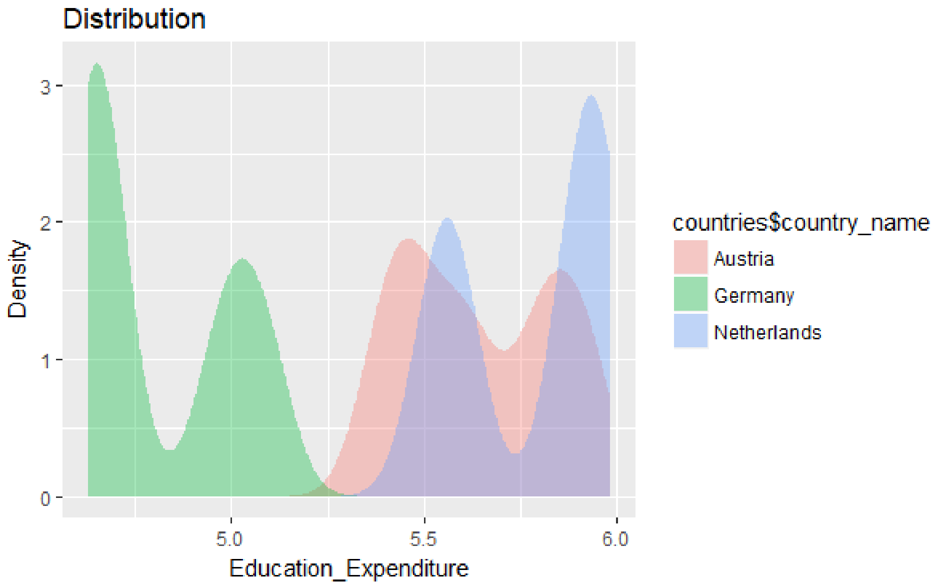


*Figure 41 Density for Countries education expenditure*

## Density Histograms for Universities dataset per variable

KPIs-> teaching, research, citations, industry_income, international_outlook

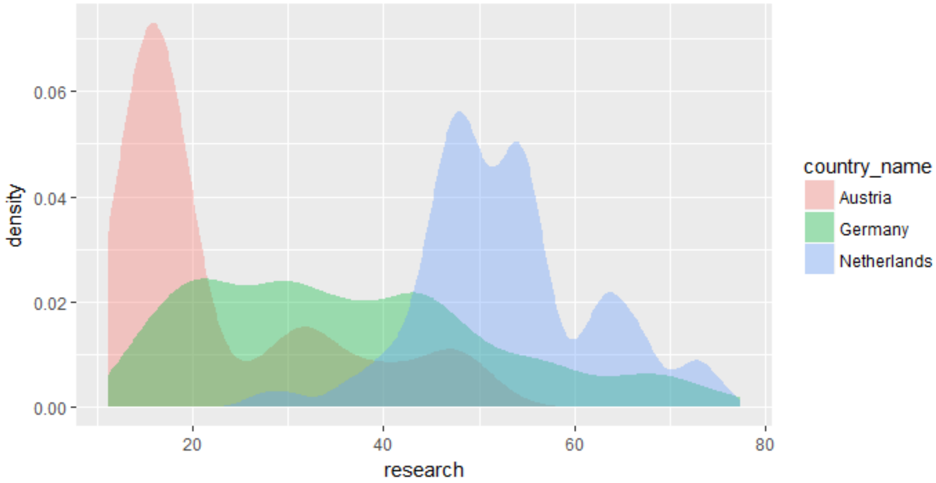


*Figure 42 Density for Countries research*

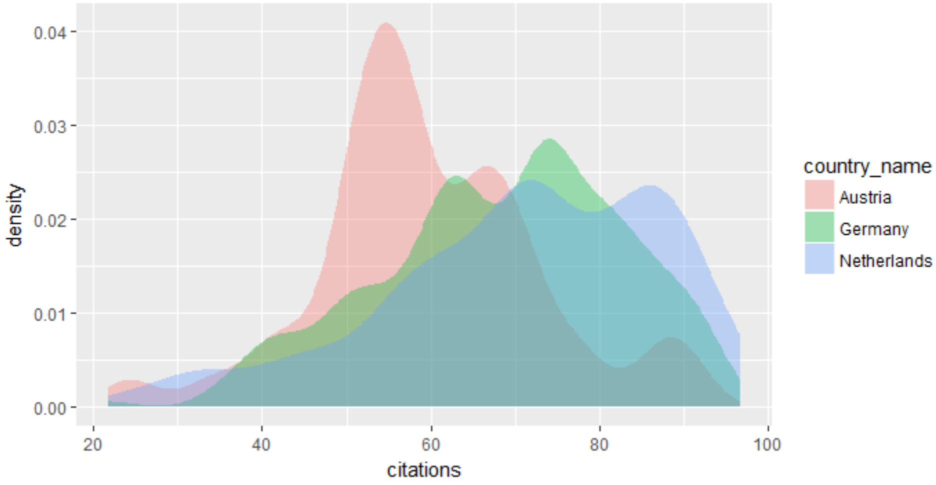


*Figure 43 Density for Countries citations*

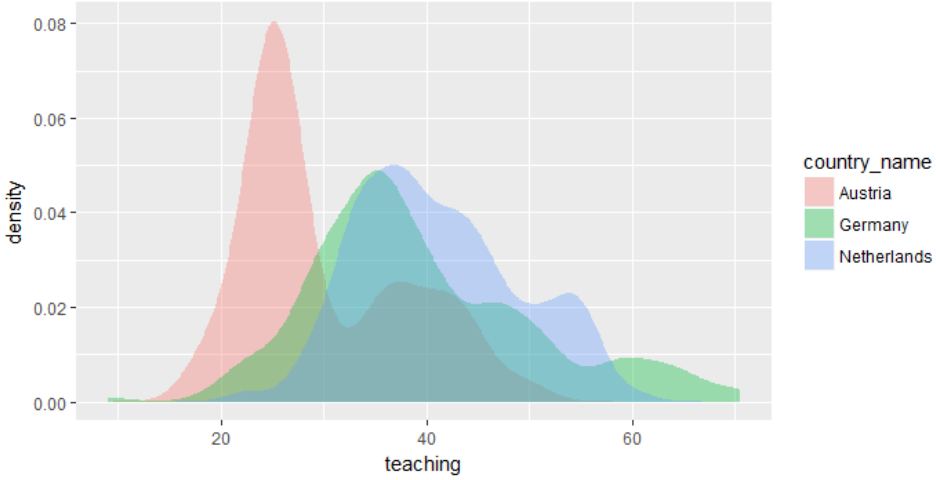


*Figure 44 Density for Countries teaching*

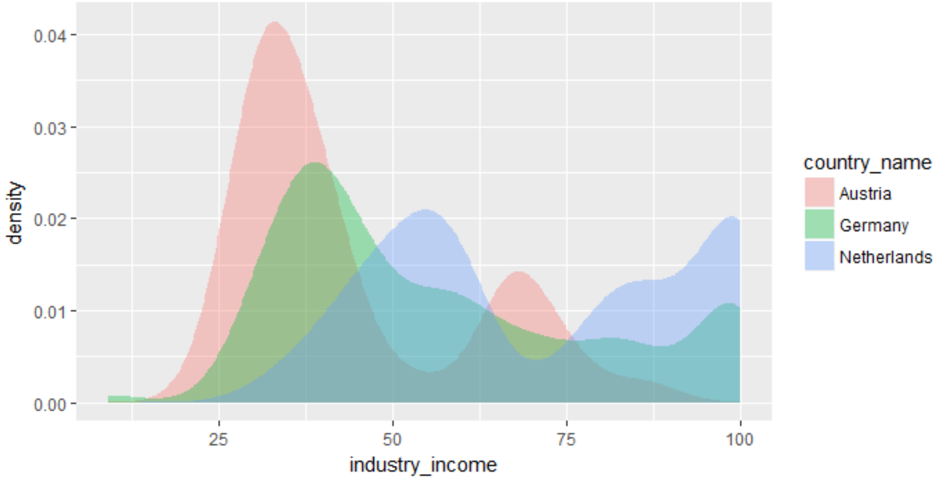


*Figure 45 Density for Countries industry income*

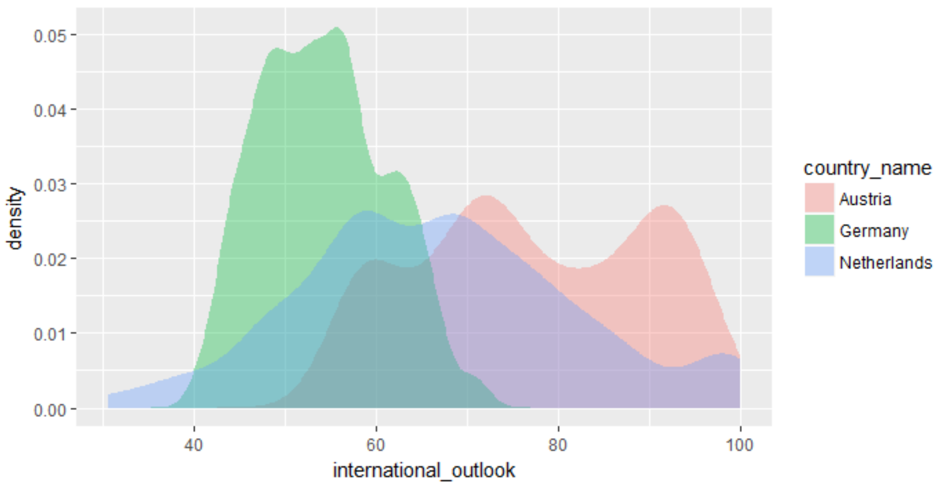
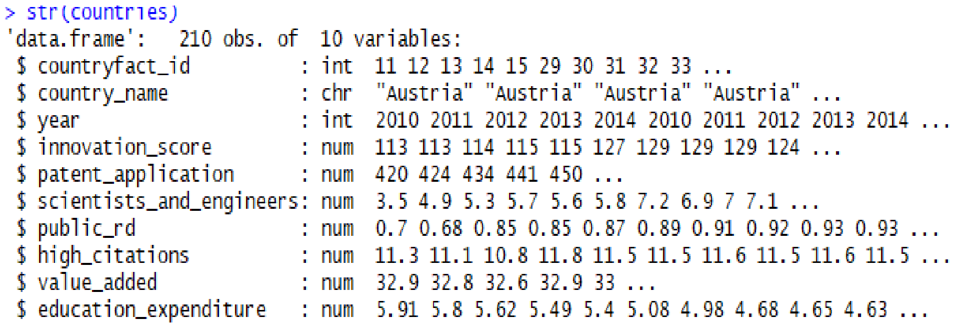


*Figure 46 Density for Countries international outlook*

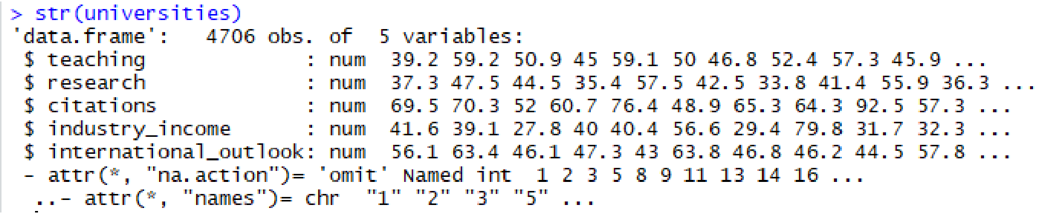
KPIs->innovation_score, patent_application, scientists_and_engineers, public_rd, hgh_citations, value_added, education_expenditure



```
> str(countries)
'data.frame':	210 obs. of  10 variables:
 $ countryfact_id          : int  11 12 13 14 15 29 30 31 32 33 ...
 $ country_name            : chr  "Austria" "Austria" "Austria" "Austria" ...
 $ year                    : int  2010 2011 2012 2013 2014 2010 2011 2012 2013 2014 ...
 $ innovation_score        : num  113 113 114 115 115 127 129 129 129 124 ...
 $ patent_application      : num  420 424 434 441 450 ...
 $ scientists_and_engineers: num  3.5 4.9 5.3 5.7 5.6 5.8 7.2 6.9 7 7.1 ...
 $ public_rd               : num  0.7 0.68 0.85 0.85 0.87 0.89 0.91 0.92 0.93 0.93 ...
 $ high_citations          : num  11.3 11.1 10.8 11.8 11.5 11.5 11.6 11.5 11.6 11.5 ...
 $ value_added             : num  32.9 32.8 32.6 32.9 33 ...
 $ education_expenditure   : num  5.91 5.8 5.62 5.49 5.4 5.08 4.98 4.68 4.65 4.63 ...
```

Figure 47 Appendix– Structure 1

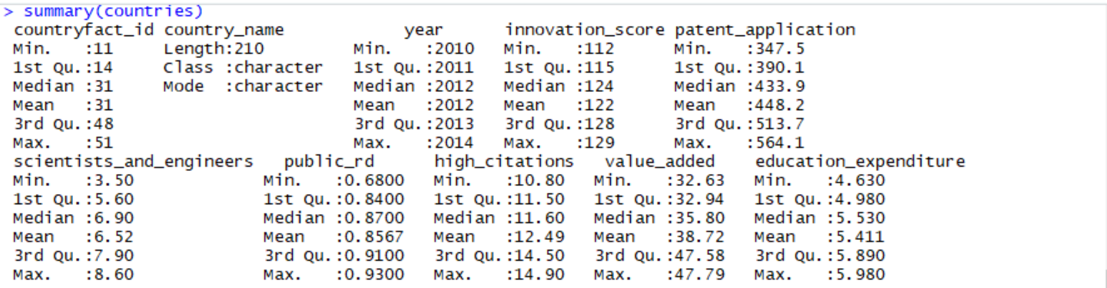
KPIs->innovation_score, patent_application, scientists_and_engineers, public_rd, hgh_citations, value_added, education_expenditure



```
> str(universities)
'data.frame':	4706 obs. of  5 variables:
 $ teaching             : num  39.2 59.2 50.9 45 59.1 50 46.8 52.4 57.3 45.9 ...
 $ research             : num  37.3 47.5 44.5 35.4 57.5 42.5 33.8 41.4 55.9 36.3 ...
 $ citations            : num  69.5 70.3 52 60.7 76.4 48.9 65.3 64.3 92.5 57.3 ...
 $ industry_income      : num  41.6 39.1 27.8 40 40.4 56.6 29.4 79.8 31.7 32.3 ...
 $ international_outlook: num  56.1 63.4 46.1 47.3 43 63.8 46.8 46.2 44.5 57.8 ...
 - attr(*, "na.action")= 'omit' Named int  1 2 3 5 8 9 11 13 14 16 ...
  ..- attr(*, "names")= chr  "1" "2" "3" "5" ...
```

Figure 48 Structure 2

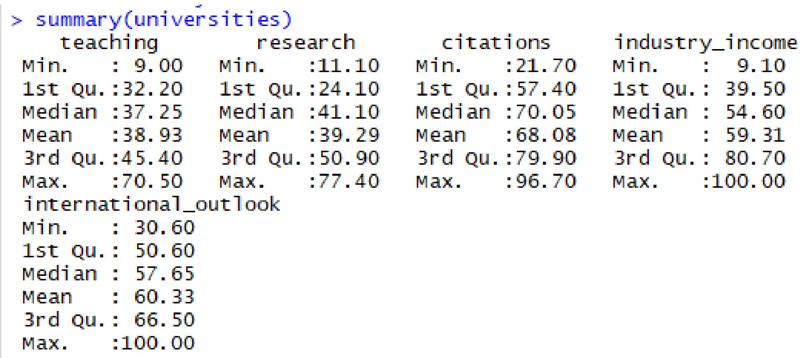
KPIs-> teaching, research, citations, industry_income, international_outlook



```
> summary(countries)
 countryfact_id country_name            year      innovation_score patent_application
 Min.   :11     Length:210         Min.   :2010   Min.   :112      Min.   :347.5     
 1st Qu.:14     Class :character   1st Qu.:2011   1st Qu.:115      1st Qu.:390.1     
 Median :31     Mode  :character   Median :2012   Median :124      Median :433.9     
 Mean   :31                        Mean   :2012   Mean   :122      Mean   :448.2     
 3rd Qu.:48                        3rd Qu.:2013   3rd Qu.:128      3rd Qu.:513.7     
 Max.   :51                        Max.   :2014   Max.   :129      Max.   :564.1     
 scientists_and_engineers   public_rd      high_citations   value_added    education_expenditure
 Min.   :3.50             Min.   :0.6800   Min.   :10.80   Min.   :32.63   Min.   :4.630        
 1st Qu.:5.60             1st Qu.:0.8400   1st Qu.:11.50   1st Qu.:32.94   1st Qu.:4.980        
 Median :6.90             Median :0.8700   Median :11.60   Median :35.80   Median :5.530        
 Mean   :6.52             Mean   :0.8567   Mean   :12.49   Mean   :38.72   Mean   :5.411        
 3rd Qu.:7.90             3rd Qu.:0.9100   3rd Qu.:14.50   3rd Qu.:47.58   3rd Qu.:5.890        
 Max.   :8.60             Max.   :0.9300   Max.   :14.90   Max.   :47.79   Max.   :5.980        
```

Figure 49 Summary 1

KPIs-> teaching, research, citations, industry_income, international_outlook



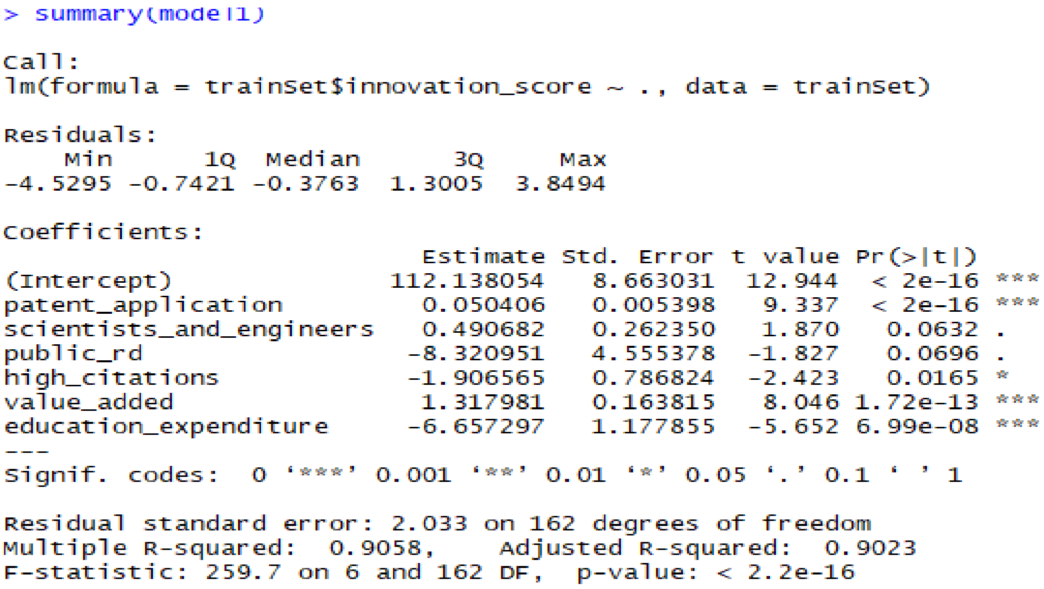
```
> summary(universities)
    teaching        research       citations      industry_income
 Min.   : 9.00   Min.   :11.10   Min.   :21.70   Min.   :  9.10
 1st Qu.:32.20   1st Qu.:24.10   1st Qu.:57.40   1st Qu.: 39.50
 Median :37.25   Median :41.10   Median :70.05   Median : 54.60
 Mean   :38.93   Mean   :39.29   Mean   :68.08   Mean   : 59.31
 3rd Qu.:45.40   3rd Qu.:50.90   3rd Qu.:79.90   3rd Qu.: 80.70
 Max.   :70.50   Max.   :77.40   Max.   :96.70   Max.   :100.00
 international_outlook
 Min.   : 30.60
 1st Qu.: 50.60
 Median : 57.65
 Mean   : 60.33
 3rd Qu.: 66.50
 Max.   :100.00
```

Figure 50 Summary 2

Predictive Model

Train Model-Countries DataSet

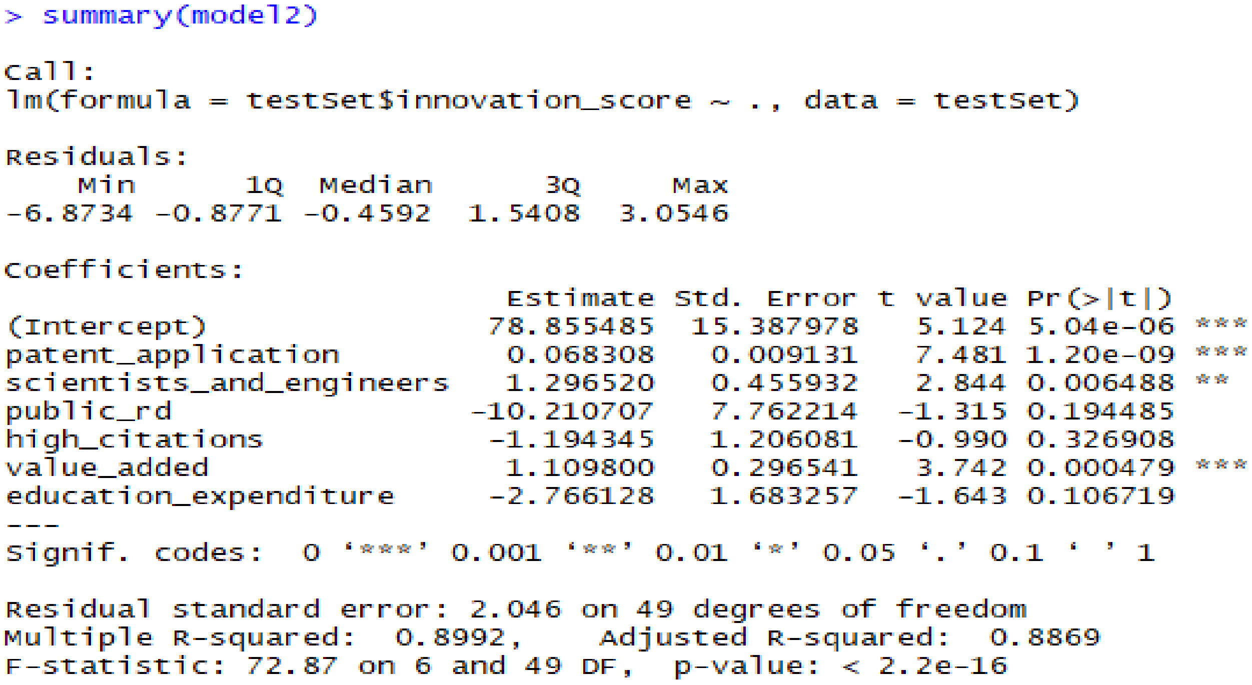
KPIs->innovation_score, patent_application, scientists_and_engineers, public_rd, hgh_citations, value_added, education_expenditure



```
> summary(model1)

Call:
lm(formula = trainSet$innovation_score ~ ., data = trainSet)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5295 -0.7421 -0.3763  1.3005  3.8494 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)              112.138054   8.663031  12.944  < 2e-16 ***
patent_application         0.050406   0.005398   9.337  < 2e-16 ***
scientists_and_engineers   0.490682   0.262350   1.870   0.0632 .  
public_rd                 -8.320951   4.555378  -1.827   0.0696 .  
high_citations            -1.906565   0.786824  -2.423   0.0165 *  
value_added                1.317981   0.163815   8.046 1.72e-13 ***
education_expenditure     -6.657297   1.177855  -5.652 6.99e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.033 on 162 degrees of freedom
Multiple R-squared:  0.9058,	Adjusted R-squared:  0.9023 
F-statistic: 259.7 on 6 and 162 DF,  p-value: < 2.2e-16
```

Figure 51 Regression Outcome TrainSet 1

Test Model-Countries TestSet

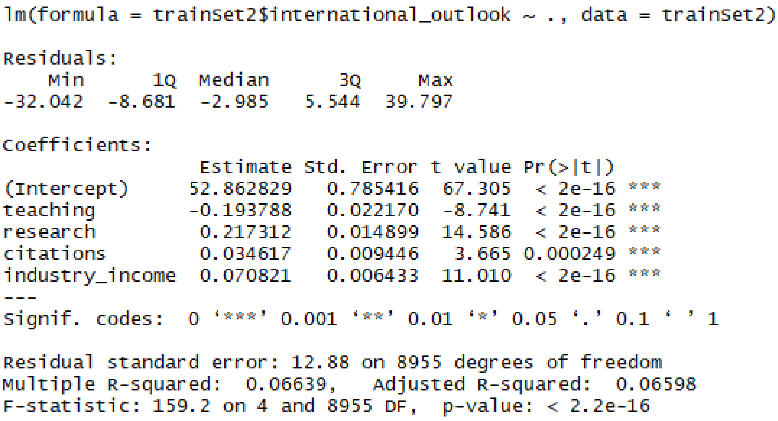


```
> summary(model2)

Call:
lm(formula = testSet$innovation_score ~ ., data = testSet)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.8734 -0.8771 -0.4592  1.5408  3.0546 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              78.855485  15.387978   5.124 5.04e-06 ***
patent_application        0.068308   0.009131   7.481 1.20e-09 ***
scientists_and_engineers  1.296520   0.455932   2.844 0.006488 ** 
public_rd               -10.210707   7.762214  -1.315 0.194485    
high_citations           -1.194345   1.206081  -0.990 0.326908    
value_added               1.109800   0.296541   3.742 0.000479 ***
education_expenditure    -2.766128   1.683257  -1.643 0.106719    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.046 on 49 degrees of freedom
Multiple R-squared:  0.8992,	Adjusted R-squared:  0.8869 
F-statistic: 72.87 on 6 and 49 DF,  p-value: < 2.2e-16
```

Figure 52 Regression Outcome TestSet

Train Model- Universities DataSet

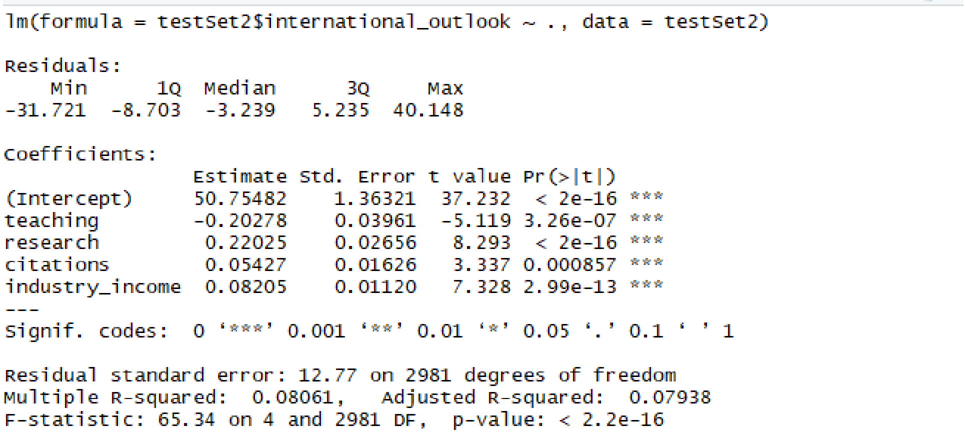
KPIs-> teaching, research, citations, industry_income, international_outlook



```
lm(formula = trainSet2$international_outlook ~ ., data = trainSet2)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.042  -8.681  -2.985   5.544  39.797 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     52.862829   0.785416  67.305  < 2e-16 ***
teaching        -0.193788   0.022170  -8.741  < 2e-16 ***
research         0.217312   0.014899  14.586  < 2e-16 ***
citations        0.034617   0.009446   3.665 0.000249 ***
industry_income  0.070821   0.006433  11.010  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.88 on 8955 degrees of freedom
Multiple R-squared:  0.06639,	Adjusted R-squared:  0.06598 
F-statistic: 159.2 on 4 and 8955 DF,  p-value: < 2.2e-16
```

Figure 53 Regression Outcome TrainSet 2

Test Model-Universities DataSet



```
lm(formula = testSet2$international_outlook ~ ., data = testSet2)

Residuals:
    Min      1Q  Median      3Q     Max 
-31.721  -8.703  -3.239   5.235  40.148 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     50.75482    1.36321  37.232  < 2e-16 ***
teaching        -0.20278    0.03961  -5.119 3.26e-07 ***
research         0.22025    0.02656   8.293  < 2e-16 ***
citations        0.05427    0.01626   3.337 0.000857 ***
industry_income  0.08205    0.01120   7.328 2.99e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.77 on 2981 degrees of freedom
Multiple R-squared:  0.08061,	Adjusted R-squared:  0.07938 
F-statistic: 65.34 on 4 and 2981 DF,  p-value: < 2.2e-16
```

Figure 54 Regression Outcome TestSet 2