 # Working paper _8_DRIVERS OF R&D EXPENDITURES . RESEARCH METHODOLOGY AND PRESENTATION OF RESULTS.

How do we prepare the research methodology correctly and how do we present the results obtained in a scientific article? I will exemplify in the context of a material related to the proposed subject.

MATERIALS AND METHODS

The research methods used in our study consist of the systematization of relevant literature (described in Section 2) and a quantitative approach based on data, statistical methods, and techniques to test the dependency hypothesis (H1), ***Digitalization, investments in Artificial Intelligence (AI), and innovation-related indicators have a significant impact on total research and development (R&D) expenditures in EU member states***.

Specifically, the empirical analysis consists of multiple regression to determine the direction of relationships and the connections between innovation determinants and R&D expenditure levels in EU countries.

Data and methodology.

- Data sources: European Commission, Eurostat, OECD.ai, World Bank.
- Analysis period: 2017-2022. Number of observations: 28 (EU member states and EU average).

Definition of the model and variables under analysis.

The **multiple regression equation** is expressed as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i \quad (1)$$

Where:

Dependent variable (*Y*):

- *RD_GDP* = Research and development (R&D) expenditures as a percentage of GDP (Source: World Bank & EUROSTAT).

Independent variables (X1÷X5):

- (X_1) *DESI* = Digital Economy and Society Index, Aggregate score, weighted score of the DESI dimension [0;100] (Source: European Commission, Digital Decade DESI visualization tool).
- (X_2) *Patents* = Patent applications to the EPO by country of inventors per one million inhabitants (Source: Eurostat).
- (X_3) *VC_invest* = Venture capital investments in AI (VC_Invest_AI), % of GDP (Source: composite indicator calculated by the authors based on OECD.ai data for venture capital investments in AI, expressed in USD millions per country, and World Bank statistics for GDP values, expressed in constant 2015 prices, USD millions).

- (X_4) *AI_projects* = AI software development. Very High-impact (>100 forks) AI projects (%) (Source: OECD.ai).
- (X_5) *Researchers* = Total number of researchers in R&D per million people (Source: World Bank, via UNESCO).

Regression coefficients: β_0 = Intercept (constant), $\beta_1 \div \beta_5$ = Regression coefficients.
ε = Residual error.
To handle predictors' missing data, we employed two approaches: calculating the average values (arithmetic mean of neighboring values) and, in certain cases, utilizing Python-based predictive models for data imputation (linear regression technique). EU-level values were calculated as weighted averages based on either population or GDP, depending on the case.
Below, we will describe the *statistical methods and techniques* used in our analysis.
The *Pearson correlation* coefficients were used to evaluate the strength of relationships between variables (theoretical range: 0–1, preferred range: 0.50–0.95). In *regression analysis*, the coefficient of determination (R^2) is crucial as it shows the percentage of variation in the dependent variable explained by the independent variables. The *statistical significance* (Sig.) should ideally be below 0.05, indicating over 95% confidence. On the other hand, a *factorial analysis* was conducted using the Kaiser-Meyer-Olkin (KMO) statistical test to assess the internal consistency of the selected variables. The KMO should range from 0.5 to 1, indicating adequate sampling. To address the issue of *multicollinearity* and obtain more robust results, we applied the following solutions in our analysis: calculating the Variance Inflation Factor (VIF) and, when necessary, implementing regularization methods. In general, a VIF below 10 (preferably below 5) does not indicate severe multicollinearity issues.

RESULTS AND DISCUSSION

Subsequently, we present the findings of our quantitative analysis for the initial and final years in the dataset (2017 and 2022). This includes the bivariate correlation analysis, regression analysis, variance inflation factor (VIF) calculation, Kaiser-Meyer-Olkin (KMO) test, and a descriptive statistical assessment for the interval endpoints, specifically 2017 and 2022.

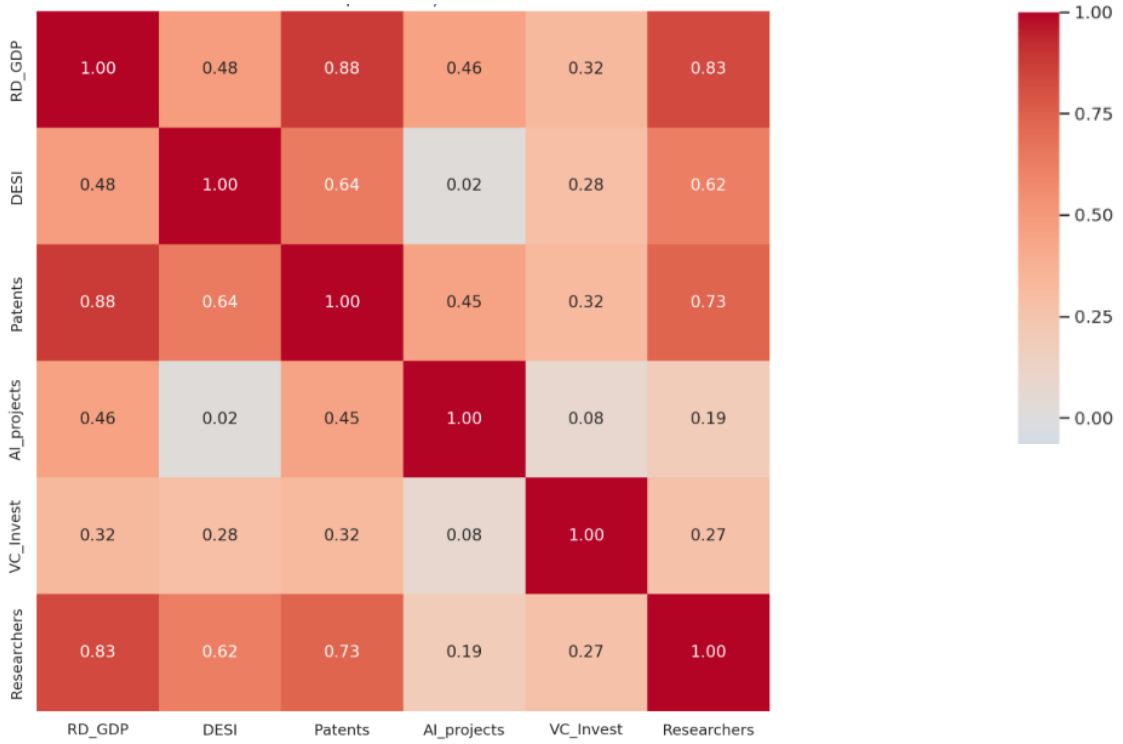

Figure 3. Correlation outputs_heatmap_2017

OLS Regression Results

Dep. Variable:	RD_GDP	R-squared:	0.899
Model:	OLS	Adj. R-squared:	0.876
Method:	Least Squares	F-statistic:	39.01
Date:	Wed, 05 Mar 2025	Prob (F-statistic):	3.23e-10
Time:	16:01:27	Log-Likelihood:	-3.5621
No. Observations:	28	AIC:	19.12
Df Residuals:	22	BIC:	27.12
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.7694	0.362	2.125	0.045	0.019	1.520
DESI	-0.0260	0.011	-2.299	0.031	-0.049	-0.003
Patents	0.0042	0.001	4.670	0.000	0.002	0.006
AI_projects	6.9599	6.088	1.143	0.265	-5.665	19.585
VC_Invest	1.7758	2.526	0.703	0.489	-3.463	7.014
Researchers	0.9454	0.193	4.893	0.000	0.545	1.346

Omnibus:	3.359	Durbin-Watson:	1.967
Prob(Omnibus):	0.187	Jarque-Bera (JB):	1.451
Skew:	0.089	Prob(JB):	0.484
Kurtosis:	1.899	Cond. No.	1.77e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.77e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Figure 4. Regression outputs_2017

Interpretation of Results 2017 (see Fig. 3 and Fig. 4):

Correlations (2017): These results show strong correlations between *R&D_GDP* and *Patents* (88%), *R&D_GDP* and *Researchers* (83%), and *Patents* and *Researchers* (73%). The correlation heatmap (Fig. 3) visually confirms the strong relationships between the previously mentioned variables.

Linear Regression (2017): The model explains about 90% of the variation in *R&D expenditures* (R-squared = 0.899). Statistically significant variables ($p < 0.05$) include *DESI* (coefficient = -0.0260), *Patents* (coefficient = 0.0042), and *Researchers* (coefficient = 0.9454). Regarding the negative *DESI* coefficient, a possible explanation is that a more digitalized society (with a higher DESI value) may promote process efficiency and investment redistribution, resulting in a lower percentage of R&D spending. This is because digital technologies can support innovation without requiring the structured costs associated with traditional investments. *AI_projects* has a coefficient of 6.9599, but it is not significant ($p = 0.265$). *VC_Invest* has a coefficient of 1.7758, but it is also not significant ($p = 0.489$). The results indicate that, in 2017, the variables that appear to significantly influence *R&D_%GDP* are DESI total, the number of patents, and the percentage of researchers. The condition number is high (1.77e+04), suggesting potential multicollinearity issues. To address the issue of multicollinearity, we calculate the Variance Inflation Factor (VIF).

Table 1. Variance Inflation Factor (VIF) value _ 2017

N	**Variable**	**Variance Inflation Factor**
0	constant	38.18
1	DESI	2.11
2	Patents	3.50
3	AI_projects	1.51
4	VC_Invest	1.13
5	Researchers	2.37

Observations: The constant has a very high VIF value (38.18), indicating a strong correlation with the other variables. All predictors have VIF values below 5, indicating low multicollinearity. Consequently, no regularization methods are required to stabilize the coefficients.

Table 2. Kaiser-Meyer-Olkin (KMO) test_2017

N	**Variable**	**Kaiser-Meyer-Olkin Value**
0	RD_GDP	0.6102
1	DESI	0.5757
2	Patents	0.6774
3	AI_projects	0.6902
4	VC_Invest	0.8747
5	Researchers	0.6639
	General	**0.6469**

KMO = 0,65 > 0.50 (50%), acceptable value.

Table 3. Descriptive statistics_2017

N	Statistic	RF_GDP	DESI	Patents	AI_projects	VC_Invest	Researchers
1	Count	28	28	28	28	28	28
2	mean	1.5882	35.023	113.5678	0.0063	0.0024	1.2385
3	std	0.8789	7.6660	125.4463	0.0120	0.0250	0.4754
4	min	0.51	19.3991	4.84	0	0	0.34
5	25%	0.895	29.6258	15.3325	0.0002	0.0046	0.88
6	50%	1.28	35.5908	44.61	0.0016	0.0178	1.245
7	75%	2.155	41.3370	182.5425	0.0061	0.0309	1.505
8	max	3.39	47.8506	370.08	0.0566	0.1114	2.09

The average *R&D expenditure* as a percentage of GDP in 2017 was approximately 1.59%, with values ranging from 0.51% to 3.39%.

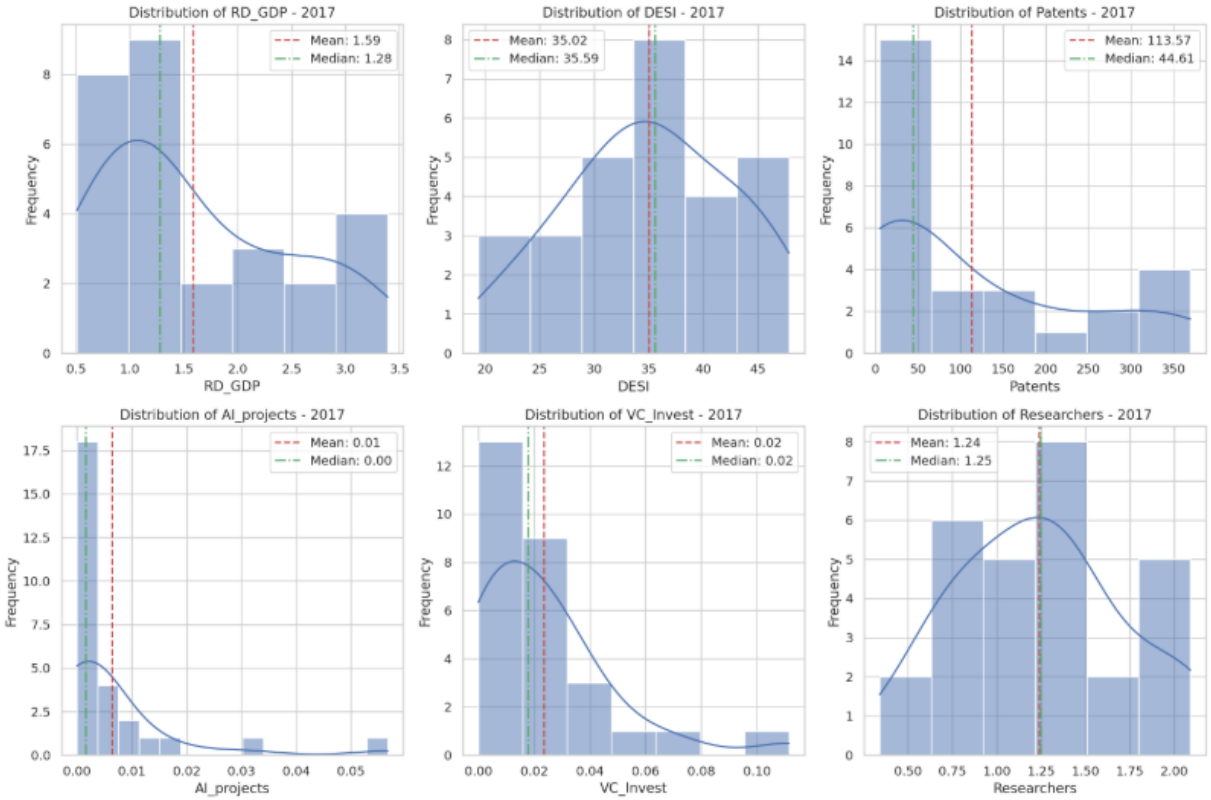

Figure 5. Histograms of Variables – 2017

The variables *Patents*, *AI_projects*, and *VC_Invest* exhibit positive skewness, indicating that most countries have low values, with only a few recording significantly higher values.

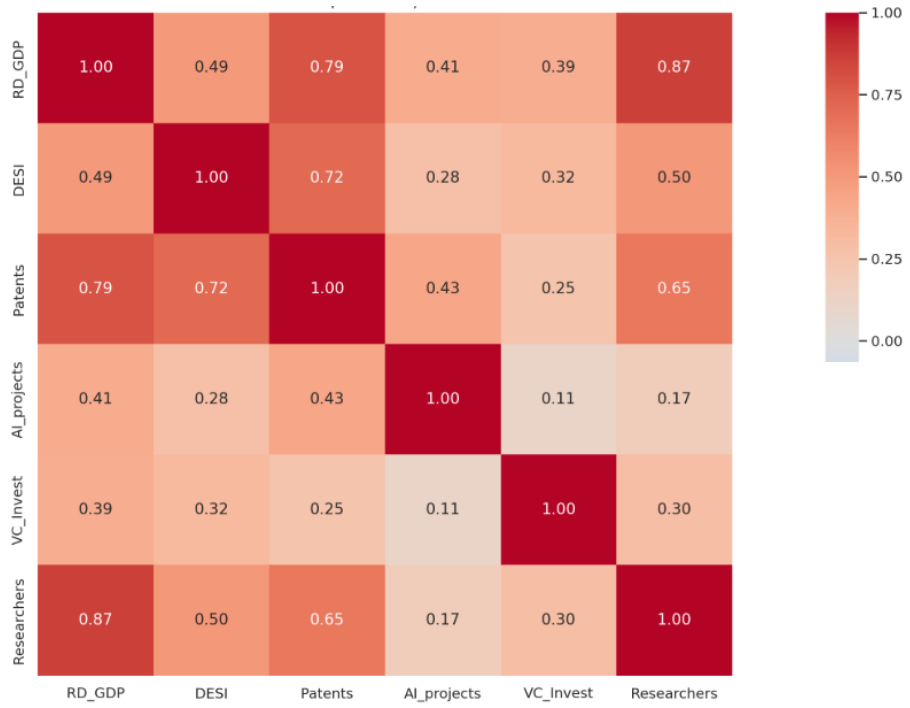

Figure 6. Correlation outputs_heatmap_2022, *Reference: authors' elaboration.*

OLS Regression Results

Dep. Variable:	RD_GDP	R-squared:	0.898
Model:	OLS	Adj. R-squared:	0.874
Method:	Least Squares	F-statistic:	38.58
Date:	Wed, 05 Mar 2025	Prob (F-statistic):	3.60e-10
Time:	15:55:32	Log-Likelihood:	-4.0547
No. Observations:	28	AIC:	20.11
Df Residuals:	22	BIC:	28.10
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8574	0.451	1.903	0.070	-0.077	1.792
DESI	-0.0208	0.009	-2.266	0.034	-0.040	-0.002
Patents	0.0030	0.001	3.663	0.001	0.001	0.005
AI_projects	18.7336	9.444	1.984	0.060	-0.852	38.319
VC_Invest	0.4063	0.196	2.069	0.050	-0.001	0.813
Researchers	0.9569	0.143	6.704	0.000	0.661	1.253

Omnibus:	1.846	Durbin-Watson:	2.207
Prob(Omnibus):	0.397	Jarque-Bera (JB):	1.137
Skew:	0.148	Prob(JB):	0.566
Kurtosis:	2.059	Cond. No.	2.96e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.96e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 7. Regression outputs_2022, *Reference: authors' elaboration.*

Interpretation of Results_2022 (see Fig. 6 and Fig. 7):

Correlations (2022): There is a strong correlation between *R&D_%GDP* and *Researchers* (86.7%). A strong correlation exists between *R&D_%GDP* and *Patents* (79%). A moderate correlation is observed between *R&D_%GDP* and *DESI* (49.3%). Weak to moderate correlations are found between *R&D_%GDP* and *AI_projects* (40.6%) and *VC_Invest* (39.2%).

Linear Regression (2022): The model explains 89.8% of the variation in *R&D expenditure as a percentage of GDP* (R-squared: 0.898). Statistically significant variables ($p < 0.01$*, $p < 0.05$**, $p < 0.1$***). All predictors exhibit statistical significance as follows:

- The negative coefficient of *DESI* (-0.0208), statistically significant with $p = 0.034$** < 0.05, suggests that an increase in a society's digitalization may influence the restructuring of budget allocations for research and development (R&D). A higher *DESI* value indicates a high level of digital technology adoption, which in certain contexts may lead to the substitution of traditional *R&D* investments with advanced technological solutions and optimized costs. Another possible explanation could be associated with the phenomenon of multicollinearity and the complex interdependencies between variables. Although the bivariate analysis shows a moderate 49% correlation between *DESI* and the share of *R&D expenditure in GDP*, this intuitively positive relationship can turn negative when other variables from the model are considered. Thus, in the presence of additional factors, an increase in *DESI* may slightly reduce the share allocated to *R&D* as other dimensions of digitalization become more relevant.
- *Patents* (positive coefficient: 0.0030*, statistically significant $p = 0.001$* < 0.01)
- *VC_Invest* (positive coefficient: 0.4063, at the significance threshold $p = 0.050$*** < 0.1)
- *Researchers* (positive coefficient: 0.9569, most significant $p = 0.000$* < 0.01)
- *AI_projects* has a large coefficient (18.7336) and is at the threshold of statistical significance ($p = 0.060$*** < 0.1)

The percentage of *researchers*, *patents*, *AI_projects*, and *VC_Invest* have a positive impact on *R&D expenditures*, while *DESI* has a negative impact.

There are potential issues with multicollinearity (Cond. No. = 2.96e+04, high value). To clarify, we calculate the Variance Inflation Factor (VIF).

Table 4. Variance Inflation Factor (VIF) value _ 2022

N	Variable	Variance Inflation Factor
0	constant	57.11
1	DESI	2.15
2	Patents	3.13
3	AI_projects	1.27
4	VC_Invest	1.15
5	Researchers	1.85

We observe that the constant has a very high VIF value (57.11), indicating a strong correlation with the other variables. All predictors have VIF values below 5, indicating low multicollinearity. As such, regularization techniques are not required.

Table 5. Kaiser-Meyer-Olkin (KMO) test_2022

N	Variable	Kaiser-Meyer-Olkin Value
0	RD_GDP	0.5549
1	DESI	0.5683
2	Patents	0.6475
3	AI_projects	0.5746
4	VC_Invest	0.4914
5	Researchers	0.5831
	General	**0.5804**

KMO = 0,58 (58%) > 0.50 (50%), acceptable value.

Table 6. Descriptive statistics_2022

N	Statistic	RF_GDP	DESI	Patents	AI_projects	VC_Invest	Researchers
1	Count	28	28	28	28	28	28
2	mean	1.7447	52.5196	126.5557	0.0048	0.2503	1.4805
3	std	0.8901	9.6790	133.4195	0.0072	0.3315	0.5780
4	min	0.46	30.5849	4.57	0	0.0114	0.36
5	25%	1.05	48.1507	23.99	0.00005	0.0545	1.105
6	50%	1.46	52.4948	54.565	0.0002	0.1276	1.4613
7	75%	2.2125	59.3317	217.3625	0.0093	0.2859	1.745
8	max	3.47	69.5976	431.81	0.02626	1.4912	2.6642

The average *R&D expenditure as a percentage of GDP* in 2022 was approximately 1.74%, with values ranging from 0.46% to 3.47%. The descriptive statistics have been visually represented through the histograms below. The histograms below visually represent the descriptive statistics.

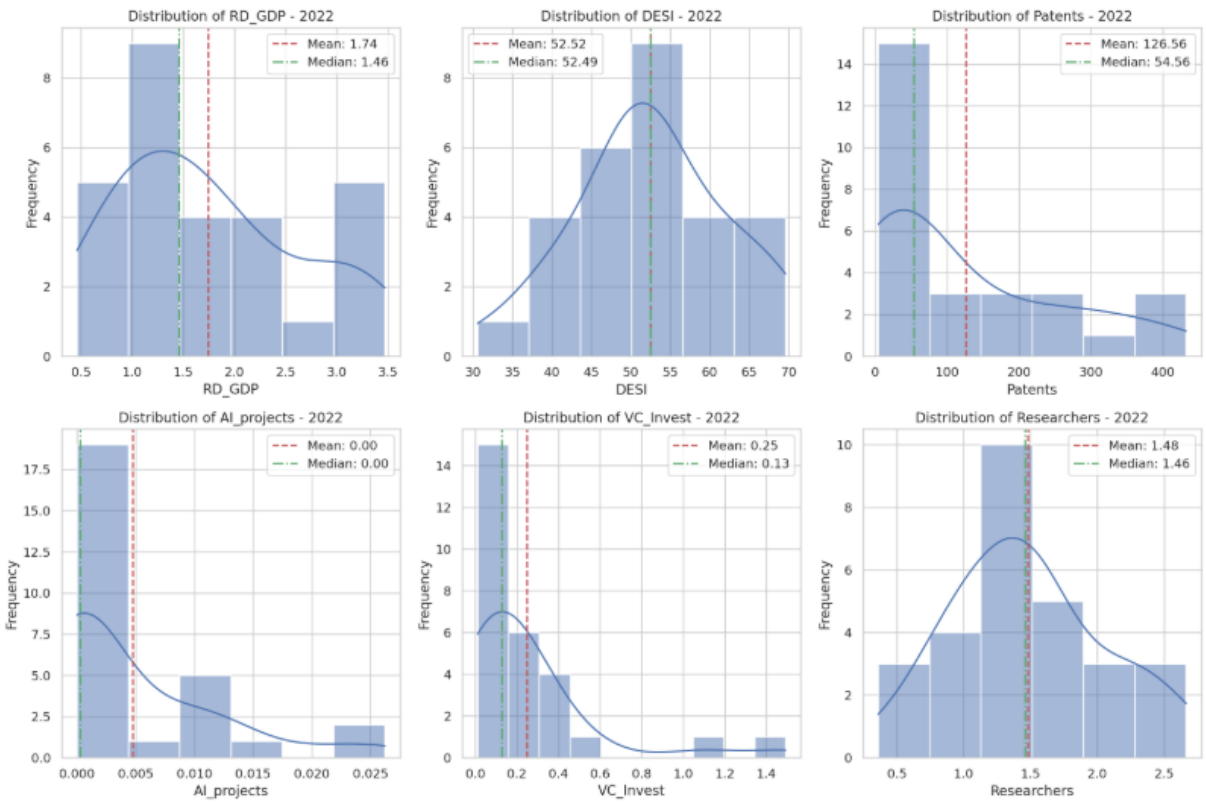

Figure 8. Histograms of Variables – 2022, *Reference: authors' elaboration.*

These histograms provide a visual representation of how each variable is distributed across countries. The red dashed lines represent the mean values, while the green dash-dot lines show the median values. You can observe several patterns:

- For both years (Fig. 5 and Fig. 8), variables like *Patents*, *AI_projects*, and *VC_Invest* show right-skewed distributions (positive skew), meaning most countries have lower values, with a few countries having significantly higher values.
- The *DESI index* shows a more symmetric distribution, especially in 2022.
- The difference between mean and median values indicates the degree of skewness in each distribution.

Comparing 2017 to 2022, you can see how the distributions have evolved.
These visualizations help identify outliers and understand the central tendency and spread of each variable across the European countries in the dataset.

Cristian – Romeo SPĂTARU, PhD candidate
Department of Economics, The Doctoral School of Economics and Business Administration,
"Alexandru Ioan Cuza" University of Iasi, Iași, Romania,

DEAR COLLEAGUES, I HOPE IT WILL HELP YOU!