Course: Data Analysis

Control Work № 1

The file **index2022\_data.xls** contains data about 184 countries.

**Variables description**

* CountryID - ID of the Country
* CountryName - Name of the Country
* HighIncome - Belonging to high income group of countries (1 - belongs, 2 - no)
* Income - Belonging to the income group (low income, middle income, high income
* Development - Is the country developed or developing
* Region - Region of the country's location
* WorldRank - World Rank
* RegionRank - Region Rank
* 2022Score - 2022 Score
* PropertyRights - Property Rights
* JudicialEffectiveness - Judicial Effectiveness
* GovernmentIntegrity - Government Integrity
* TaxBurden - Tax Burden
* GovernmentSpending - Government Spending
* FiscalHealth - Fiscal Health
* BusinessFreedom - Business Freedom
* LaborFreedom - Labor Freedom
* MonetaryFreedom - Monetary Freedom
* TradeFreedom - Trade Freedom
* InvestmentFreedom - Investment Freedom
* FinancialFreedom - Financial Freedom
* TariffRate - Tariff Rate (%)
* IncomeTaxRate - Income Tax Rate (%)
* CorporateTaxRate - Corporate Tax Rate (%)
* TaxBurdenofGDP - Tax Burden % of GDP
* GovernmentExpenditurofGDP - Gov't Expenditure % of GDP
* Population - Population (Millions)
* GDP - GDP (Billions, PPP)
* GDP\_GrowthRate - GDP Growth Rate (%)
* 5Year\_GDP\_GrowthRate - 5 Year GDP Growth Rate (%)
* GDPperCapita - GDP per Capita (PPP)
* Unemployment - Unemployment (%)
* Inflation - Inflation (%)
* FDIInflow - FDI Inflow (Millions)
* PublicDebt - Public Debt (% of GDP)

| **Student** | **Linear regression: Dependent variable** | **Linear regression: Predictors** | **Cluster analysis** |
| --- | --- | --- | --- |
| Абросов Сергей Андреевич | PropertyRights | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with PropertyRights | PropertyRights + not less than 5 continuous variables which have the correlation with PropertyRights less than 0,5 |
| Анисин Александр Александрович | JudicialEffectiveness | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with JudicialEffectiveness | JudicialEffectiveness + not less than 5 continuous variables which have the correlation with JudicialEffectiveness less than 0,5 |
| Аношенко Даниил Сергеевич | GovernmentIntegrity | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with GovernmentIntegrity | GovernmentIntegrity + not less than 5 continuous variables which have the correlation with GovernmentIntegrity less than 0,5 |
| Верховод Никита Сергеевич | TaxBurden | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with TaxBurden | TaxBurden + not less than 5 continuous variables which have the correlation with TaxBurden less than 0,5 |
| Дандамаев Гаджи - | GovernmentSpending | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with GovernmentSpending | GovernmentSpending + not less than 5 continuous variables which have the correlation with GovernmentSpending less than 0,5 |
| Дерезовский Илья Денисович | BusinessFreedom | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with BusinessFreedom | BusinessFreedom + not less than 5 continuous variables which have the correlation with BusinessFreedom less than 0,5 |
| Дроздова Ксения Александровна | LaborFreedom | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with LaborFreedom | LaborFreedom + not less than 5 continuous variables which have the correlation with LaborFreedom less than 0,5 |
| Дубинский Роман Вячеславович | MonetaryFreedom | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with MonetaryFreedom | MonetaryFreedom + not less than 5 continuous variables which have the correlation with MonetaryFreedom less than 0,5 |
| Думкин Никита Алексеевич | TradeFreedom | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with TradeFreedom | TradeFreedom + not less than 5 continuous variables which have the correlation with TradeFreedom less than 0,5 |
| Жулин Артем - | InvestmentFreedom | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with InvestmentFreedom | InvestmentFreedom + not less than 5 continuous variables which have the correlation with InvestmentFreedom less than 0,5 |
| Кара Дмитрий Александрович | FinancialFreedom | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with FinancialFreedom | FinancialFreedom + not less than 5 continuous variables which have the correlation with FinancialFreedom less than 0,5 |
| Караваева Екатерина Андреевна | TariffRate | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with TariffRate | TariffRate + not less than 5 continuous variables which have the correlation with TariffRate less than 0,5 |
| Кочнев Виктор Алексеевич | IncomeTaxRate | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with IncomeTaxRate | IncomeTaxRate + not less than 5 continuous variables which have the correlation with IncomeTaxRate less than 0,5 |
| Кузьменко Роман Олегович | CorporateTaxRate | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with CorporateTaxRate | CorporateTaxRate + not less than 5 continuous variables which have the correlation with CorporateTaxRate less than 0,5 |
| Кузьмин Максим Антонович | TaxBurdenofGDP | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with TaxBurdenofGDP | TaxBurdenofGDP + not less than 5 continuous variables which have the correlation with TaxBurdenofGDP less than 0,5 |
| Ланин Георгий Михайлович | GovernmentExpenditurofGDP | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with GovernmentExpenditurofGDP | GovernmentExpenditurofGDP + not less than 5 continuous variables which have the correlation with GovernmentExpenditurofGDP less than 0,5 |
| Лохматиков Сергей Павлович | Inflation | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with Inflation | Inflation + not less than 5 continuous variables which have the correlation with Inflation less than 0,5 |
| Лукашин Даниил Дмитриевич | GDP\_GrowthRate | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with GDP\_GrowthRate | GDP\_GrowthRate + not less than 5 continuous variables which have the correlation with GDP\_GrowthRate less than 0,5 |
| Малыш Игорь Игоревич | 5Year\_GDP\_GrowthRate | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with 5Year\_GDP\_GrowthRate | 5Year\_GDP\_GrowthRate + not less than 5 continuous variables which have the correlation with 5Year\_GDP\_GrowthRate less than 0,5 |
| Микулик Даниил Павлович | Unemployment | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with Unemployment | Unemployment + not less than 5 continuous variables which have the correlation with Unemployment less than 0,5 |
| Нгуен Дык Хю - | GDPperCapita | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with GDPperCapita | GDPperCapita + not less than 5 continuous variables which have the correlation with GDPperCapita less than 0,5 |
| Пань Чжэну - | Inflation | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with Inflation | Inflation + not less than 5 continuous variables which have the correlation with Inflation less than 0,5 |
| Правира Денуне Галанг - | FDIInflow | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with FDIInflow | FDIInflow + not less than 5 continuous variables which have the correlation with FDIInflow less than 0,5 |
| Рогачев Сергей Алексеевич | PublicDebt | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with PublicDebt | PublicDebt + not less than 5 continuous variables which have the correlation with PublicDebt less than 0,5 |
| Рогоза Кирилл Андреевич | GDP | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with GDP | GDP + not less than 5 continuous variables which have the correlation with GDP less than 0,5 |
| Сергеев Илья Иванович | FiscalHealth | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with FiscalHealth | FiscalHealth + not less than 5 continuous variables which have the correlation with FiscalHealth less than 0,5 |
| Сигал Лев Золтанович | JudicialEffectiveness | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with JudicialEffectiveness | JudicialEffectiveness + not less than 5 continuous variables which have the correlation with JudicialEffectiveness less than 0,5 |
| Сидоренков Олег Владимирович | PublicDebt | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with PublicDebt | PublicDebt + not less than 5 continuous variables which have the correlation with PublicDebt less than 0,5 |
| Смолкин Михаил Романович | GDP\_GrowthRate | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with GDP\_GrowthRate | GDP\_GrowthRate + not less than 5 continuous variables which have the correlation with GDP\_GrowthRate less than 0,5 |
| Солозобов Евгений Дмитриевич | PropertyRights | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with PropertyRights | PropertyRights + not less than 5 continuous variables which have the correlation with PropertyRights less than 0,5 |
| Терло Адам Зурабович | JudicialEffectiveness | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with JudicialEffectiveness | JudicialEffectiveness + not less than 5 continuous variables which have the correlation with JudicialEffectiveness less than 0,5 |
| Тимонин Антон Сергеевич | GovernmentIntegrity | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with GovernmentIntegrity | GovernmentIntegrity + not less than 5 continuous variables which have the correlation with GovernmentIntegrity less than 0,5 |
| Тимонин Егор Сергеевич | TaxBurden | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with TaxBurden | TaxBurden + not less than 5 continuous variables which have the correlation with TaxBurden less than 0,5 |
| Тюкавкина Екатерина Николаевна | GovernmentSpending | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with GovernmentSpending | GovernmentSpending + not less than 5 continuous variables which have the correlation with GovernmentSpending less than 0,5 |
| Фам Тиен Мань - | BusinessFreedom | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with BusinessFreedom | BusinessFreedom + not less than 5 continuous variables which have the correlation with BusinessFreedom less than 0,5 |
| Цатурьян Константин Артурович | LaborFreedom | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with LaborFreedom | LaborFreedom + not less than 5 continuous variables which have the correlation with LaborFreedom less than 0,5 |
| Цветкова Мария Андреевна | MonetaryFreedom | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with MonetaryFreedom | MonetaryFreedom + not less than 5 continuous variables which have the correlation with MonetaryFreedom less than 0,5 |
| Цуркан Даниел Валентинович | TradeFreedom | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with TradeFreedom | TradeFreedom + not less than 5 continuous variables which have the correlation with TradeFreedom less than 0,5 |
| Штанов Андрей Николаевич | InvestmentFreedom | Income, Development, Region + not less than 5 continuous  variables which have the highest correlation with InvestmentFreedom | InvestmentFreedom + not less than 5 continuous variables which have the correlation with InvestmentFreedom less than 0,5 |
| Беловицкий Владислав | PropertyRights | HighIncome, Development, Region + not less than 5 continuous  variables which have the highest correlation with PropertyRights | PropertyRights + not less than 5 continuous variables which have the correlation with PropertyRights less than 0,5 |

Please select variables for the analysis based on the information from the above table and consider the following tasks.

**Part 1**

* 1. Do the descriptive analysis of all the variables which will be used in cluster analysis. Prove that the selected variables could be used simultaneously in one cluster model.
  2. Select a suitable cluster algorithm. Explain your selection.
  3. Define the number of clusters. Explain your decision.
  4. Describe the clusters. Each cluster should have its profile: name and detailed description in textual form based on the calculated statistical characteristic. Explain the selection of names.

**Part 2**

* 1. Do the descriptive analysis of all the variables which will be used in regression analysis. Prove that the selected predictors could be used simultaneously in one regression model.
  2. Create a linear regression model. Formulate the research hypotheses which will be verified during the regression analysis.
  3. Specify the regression equation. Assess the goodness-of-fit of the model.
  4. Provide a detailed interpretation of the relationship between each predictor and the dependent variable.
  5. Perform the model diagnostics.
  6. Make final conclusions about the results of regression analysis.

|  |  |
| --- | --- |
| **Point of the task** | **Weight of each point** |
| 1.1. | 8 % |
| 1.2. | 6 % |
| 1.3. | 8 % |
| 1.4. | 20 % |
| 2.1. | 8 % |
| 2.2. | 8 % |
| 2.3. | 8 % |
| 2.4. | 20 % |
| 2.5. | 8 % |
| 2.6. | 6 % |
| *Total* | *100 %* |

Please upload the Jupyter Notebook with comments and answers in Smart LMS. The deadline is 18 October 14:15. The task should be done **individually**.