Научно-исследовательский семинар «Анализ данных в Python»

Контрольная работа №1

В файле **index2022\_data.xls** содержаться данные, характеризующие различные показатели деятельности стран мира за 2022 год.

**Описание переменных**

* CountryID – Идентификатор страны
* CountryName – Название страны
* HighIncome – Принадлежность к странам с высоким доходом (1 - принадлежит, 2 - нет)
* Income – Принадлежность к одной из трех групп стран по уровню дохода (low income, middle income, high income)
* Development – Является ли страна развитой или развивающейся (developing, developed)
* Region – регион, в котором расположена страна
* PropertyRights - Индекс защиты прав собственности (чем эффективнее правовая защита собственности, тем выше значение переменной)
* JudicialEffectiveness – Индекс эффективности правовой системы (чем эффективнее правовая система, тем выше значение переменной)
* GovernmentIntegrity – Индекс честности правительства (чем менее коррумпировано правительство, тем выше значение переменной)
* TaxBurden – Индекс налоговой нагрузки (чем ниже налоговая нагрузка в стране, тем выше значение переменной)
* BusinessFreedom – Индекс свободы ведения бизнеса (чем в большой степени нормативно-правовая и инфраструктурная среда страны благоприятствует эффективной работе бизнеса, тем выше значение переменной)
* LaborFreedom - Индекс свободы труда (чем лучше условия труда в стране, тем выше значение переменной)
* TradeFreedom – Индекс свободы торговли (чем меньше в стране барьеров, затрудняющих импорт и экспорт товаров и услуг, тем выше значение переменной)
* InvestmentFreedom – Индекс свободы инвестиций (чем меньше в стране ограничений на приток инвестиционного капитала, тем выше значение переменной)
* FinancialFreedom – Индекс финансовой свободы (чем эффективней работает банковская система в стране и чем более независим финансовый сектор от государственного контроля и вмешательства, тем выше значение переменной)
* TariffRate – Тарифная ставка (в %)
* IncomeTaxRate – Ставка подоходного налога (в %)
* CorporateTaxRate – Ставка корпоративного налога (в %)
* TaxBurdenofGDP – Налоговое бремя в % от ВВП
* GovernmentExpenditurofGDP – Расходы правительства в процентах от ВВП
* Population – Численность населения страны (в миллионах людей)
* GDP – ВВП страны в ее собственной валюте, поделенный на обменный курс по паритету покупательной способности (в миллиардах денежных единиц)
* GDP\_GrowthRate – Темпы роста ВВП (в %)
* 5Year\_GDP\_GrowthRate – Темпы роста ВВП за 5 лет (в %)
* GDPperCapita – ВВП страны на душу населения в ее собственной валюте, поделенный на обменный курс по паритету покупательной способности
* Unemployment – Уровень безработицы (в %)
* Inflation – Уровень инфляции (в %)
* FDIInflow – Приток прямых инвестиций (в миллионах долларов)
* PublicDebt – Государственный долг (в % от ВВП)

|  |  |  |  |
| --- | --- | --- | --- |
| **Студент** | **Линейная регрессия: зависимая переменная** | **Линейная регрессия: независимые переменные** | **Переменные для проведения кластерного анализа** |
| Агаркова Полина Александровна | PropertyRights | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной PropertyRights | PropertyRights + не менее 5 метрических переменных, корреляция которых с переменной PropertyRights, не превышает 0,5 |
| Альтшулер Эдуард Дмитриевич | JudicialEffectiveness | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной JudicialEffectiveness | JudicialEffectiveness + не менее 5 метрических переменных, корреляция которых с переменной JudicialEffectiveness, не превышает 0,5 |
| Асатуров Максим Гарриевич | GovernmentIntegrity | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной GovernmentIntegrity | GovernmentIntegrity + не менее 5 метрических переменных, корреляция которых с переменной GovernmentIntegrity, не превышает 0,5 |
| Бакланов Егор Михайлович | TaxBurden | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной TaxBurden | TaxBurden + не менее 5 метрических переменных, корреляция которых с переменной TaxBurden, не превышает 0,5 |
| Балдина Анастасия Сергеевна | BusinessFreedom | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной BusinessFreedom | BusinessFreedom + не менее 5 метрических переменных, корреляция которых с переменной BusinessFreedom, не превышает 0,5 |
| Бердинских Никита Эдуардович | LaborFreedom | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной LaborFreedom | LaborFreedom + не менее 5 метрических переменных, корреляция которых с переменной LaborFreedom, не превышает 0,5 |
| Бершицкий Дмитрий Александрович | TradeFreedom | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной TradeFreedom | TradeFreedom + не менее 5 метрических переменных, корреляция которых с переменной TradeFreedom, не превышает 0,5 |
| Бобышкин Иван Михайлович | InvestmentFreedom | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной InvestmentFreedom | InvestmentFreedom + не менее 5 метрических переменных, корреляция которых с переменной InvestmentFreedom, не превышает 0,5 |
| Бойко Евгений Романович | FinancialFreedom | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной FinancialFreedom | FinancialFreedom + не менее 5 метрических переменных, корреляция которых с переменной FinancialFreedom, не превышает 0,5 |
| Бородулин Арсений Михайлович | TariffRate | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной TariffRate | TariffRate + не менее 5 метрических переменных, корреляция которых с переменной TariffRate, не превышает 0,5 |
| Бузуруков Саидазамхон Акобирхонович | IncomeTaxRate | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной IncomeTaxRate | IncomeTaxRate + не менее 5 метрических переменных, корреляция которых с переменной IncomeTaxRate, не превышает 0,5 |
| Городецкий Егор Сергеевич | CorporateTaxRate | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной CorporateTaxRate | CorporateTaxRate + не менее 5 метрических переменных, корреляция которых с переменной CorporateTaxRate, не превышает 0,5 |
| Кажкаримов Асхат Айнабекович | TaxBurdenofGDP | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной TaxBurdenofGDP | TaxBurdenofGDP + не менее 5 метрических переменных, корреляция которых с переменной TaxBurdenofGDP, не превышает 0,5 |
| Кокорев Артём Вячеславович | GovernmentExpenditurofGDP | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной GovernmentExpenditurofGDP | GovernmentExpenditurofGDP + не менее 5 метрических переменных, корреляция которых с переменной GovernmentExpenditurofGDP, не превышает 0,5 |
| Лукин Илья Максимович | Inflation | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной Inflation | Inflation + не менее 5 метрических переменных, корреляция которых с переменной Inflation, не превышает 0,5 |
| Можеитов Матвей Андреевич | GDP\_GrowthRate | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной GDP\_GrowthRate | GDP\_GrowthRate + не менее 5 метрических переменных, корреляция которых с переменной GDP\_GrowthRate, не превышает 0,5 |
| Нгуен Туан Минь | 5Year\_GDP\_GrowthRate | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной 5Year\_GDP\_GrowthRate | 5Year\_GDP\_GrowthRate + не менее 5 метрических переменных, корреляция которых с переменной 5Year\_GDP\_GrowthRate, не превышает 0,5 |
| Орешкина Наталия Максимовна | Unemployment | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной Unemployment | Unemployment + не менее 5 метрических переменных, корреляция которых с переменной Unemployment, не превышает 0,5 |
| Павлов Виктор Вадимович | GDPperCapita | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной GDPperCapita | GDPperCapita + не менее 5 метрических переменных, корреляция которых с переменной GDPperCapita, не превышает 0,5 |
| Пасынков Матвей Евгеньевич | Inflation | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной Inflation | Population + не менее 5 метрических переменных, корреляция которых с переменной Population, не превышает 0,5 |
| Песчанский Дмитрий Александрович | FDIInflow | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной FDIInflow | FDIInflow + не менее 5 метрических переменных, корреляция которых с переменной FDIInflow, не превышает 0,5 |
| Петров Константин Сергеевич | PublicDebt | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной PublicDebt | PublicDebt + не менее 5 метрических переменных, корреляция которых с переменной PublicDebt, не превышает 0,5 |
| Полканов Кирилл | GDP | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной GDP | GDP + не менее 5 метрических переменных, корреляция которых с переменной GDP, не превышает 0,5 |
| Пономарев Николай Юрьевич | TradeFreedom | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной TradeFreedom | TradeFreedom + не менее 5 метрических переменных, корреляция которых с переменной TradeFreedom, не превышает 0,5 |
| Сёмкин Арсений Сергеевич | 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 |
| Скляр Наталья Романовна | 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 |
| Старченков Вадим Андреевич | 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 |
| Соколов Артём Михайлович | 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 |
| Тепляков Владислав Витальевич | PropertyRights | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной PropertyRights | PropertyRights + не менее 5 метрических переменных, корреляция которых с переменной PropertyRights, не превышает 0,5 |
| Хадзакос Николай Александрович | BusinessFreedom | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной BusinessFreedom | BusinessFreedom + не менее 5 метрических переменных, корреляция которых с переменной BusinessFreedom, не превышает 0,5 |
| Цимбалов Максим Андреевич | LaborFreedom | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной LaborFreedom | LaborFreedom + не менее 5 метрических переменных, корреляция которых с переменной LaborFreedom, не превышает 0,5 |
| Чапурина Валерия Сергеевна | InvestmentFreedom | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной InvestmentFreedom | InvestmentFreedom + не менее 5 метрических переменных, корреляция которых с переменной InvestmentFreedom, не превышает 0,5 |
| Шикалов Андрей Артемович | FinancialFreedom | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной FinancialFreedom | FinancialFreedom + не менее 5 метрических переменных, корреляция которых с переменной FinancialFreedom, не превышает 0,5 |
| Щуплова Анна Игоревна | TariffRate | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной TariffRate | TariffRate + не менее 5 метрических переменных, корреляция которых с переменной TariffRate, не превышает 0,5 |
| Щур Виктория Олеговна | IncomeTaxRate | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной IncomeTaxRate | IncomeTaxRate + не менее 5 метрических переменных, корреляция которых с переменной IncomeTaxRate, не превышает 0,5 |
| Кошкин Георгий Андреевич | CorporateTaxRate | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной CorporateTaxRate | CorporateTaxRate + не менее 5 метрических переменных, корреляция которых с переменной CorporateTaxRate, не превышает 0,5 |
| Лебедев Андрей Андреевич | TaxBurdenofGDP | Income, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной TaxBurdenofGDP | TaxBurdenofGDP + не менее 5 метрических переменных, корреляция которых с переменной TaxBurdenofGDP, не превышает 0,5 |
| Анисевич Егор Артемович | GovernmentExpenditurofGDP | HighIncome, Development, Region + не менее 5 метрических переменных, имеющих самую высокую корреляцию с переменной GovernmentExpenditurofGDP | GovernmentExpenditurofGDP + не менее 5 метрических переменных, корреляция которых с переменной GovernmentExpenditurofGDP, не превышает 0,5 |

Найдите строку со своей фамилией в вышеприведенной таблице и используйте переменные для анализа в соответствии с рекомендациями.

**Задание 1 Линейная регрессия**

* Проведите описательный анализ переменных, которые будут использованы при построении модели линейной регрессии.
* Проанализируйте взаимосвязи между всеми парами переменных, которые будут использованы при построении модели регрессии. Обоснуйте выбор методов исследования взаимосвязей, сформулируйте содержательные выводы по итогам проведенных расчётов.
* Сформулируйте гипотезы исследования, которые будут проверены в ходе регрессионного анализа данных.
* Постройте модель множественной линейной регрессии.
* Напишите уравнение регрессии. Оцените характеристики регрессионной модели в целом.
* Оцените статистическую значимость регрессионных коэффициентов. Сравните их между собой по силе взаимосвязи с зависимой переменной.
* Подробно интерпретируйте особенности взаимосвязей между каждым из предикторов и зависимой переменной.
* Проведите диагностику модели. Есть ли в модели выбросы и/или влияющие наблюдения? Чем это может объясняться?
* Сформулируйте общие содержательные выводы по результатам регрессионного анализа данных. Какие результаты анализа данных для вас оказались очевидными, а какие неожиданными?

**Задание 2 Кластерный анализ**

* Проведите описательный анализ всех переменных, которые будут использованы при проведении кластерного анализа данных.
* Выберите подходящий алгоритм кластеризации. Обоснуйте свой выбор.
* Определите число кластеров, объясните своё решение.
* Опишите получившиеся кластеры на основе описательной статистки. Нужно не просто рассчитать статистики в таблицах, но детально описать полученные результаты.
* Дайте названия кластерам и обоснуйте их.
* Как вы считаете, почему получилась именно такая кластерная структура? Полученные результаты соответствуют вашем изначальным предположениям или оказались неожиданными?

**Решения нужно оформить в Jupyter Notebook и загрузить в Smart LMS до 21:00 3 декабря. Ноутбук должен содержать выполненный код с результатами расчётов и текстовые блоки с комментариями. Задание выполняется каждым студентом индивидуально.**