Trabalho Prático

Análise de Dados em Informática

Engenharia Informática - 3º ano 2º semestre
Ano Letivo 2023/2024

---

1. Objetivos

Objetivo Geral:

- Análise Exploratórias de Dados
- Análise Inferencial
- Correlação e Regressão

Objetivos específicos:

- Definir a metodologia de trabalho
- Análise e discussão dos resultados com recurso ao Phyton
- Escrita de Artigo Técnico com a Análise de Dados

2. Calendarização

Lançamento das propostas de trabalhos: até 8 de março de 2024
Entrega do trabalho: até **7 de abril de 2024** (23:55)
Defesa e discussão: em data a marcar pelo professor de TP

3. Normas

- O grupo (**máx 3 elementos**) deve ser o mesmo nas 2 iterações do Trabalho Prático.
- Deverá ser usado o Phyton como ferramenta de suporte ao tratamento de dados.
- A **data final de ENTREGA** do 1º Trabalho Prático é **7 de abril de 2024**, no moodle. Independentemente deste prazo, os grupos deverão ser capazes de, quando o professor o solicitar, reportar o estado de desenvolvimento do trabalho.
- A entrega do trabalho consta de um Artigo Científico. Deverá submeter todos os documentos num ficheiro compactado. O zip file deve conter:
    - artigo científico em pdf
    - dados utilizados em formato csv
    - script completo (e comentado) do código criado em Phyton para resolver o problema proposto
- O nome do ficheiro deverá seguir a seguinte notação:

  ANADI_YYY_XXX_Nºaluno1_Nºaluno2_Nºaluno3.zip, onde **YYY** representa a sigla do docente das TP, e **XXX** representa a turma TP.

  Exemplo: **ANADI_AIM_3DA_7777777_8888888_9999999.zip.**
- Trabalhos cuja designação não respeite a notação indicada, **serão penalizados em 10%.**
- **A entrega do trabalho deverá ser submetida no moodle até à data de entrega definida. Não serão aceites trabalhos fora do prazo.**
- A defesa e discussão dos trabalhos decorrerá em dia e hora a marcar por cada professor das teórico-práticas. No dia da apresentação, **TODOS** os elementos do grupo deverão estar presentes. Os elementos ausentes não terão classificação. A defesa e discussão serão realizadas em grupo com questões direcionadas a cada elemento individualmente.
- Cada grupo é responsável por gerir o seu processo de desenvolvimento. Dificuldades e problemas deverão ser comunicados atempadamente ao professor das aulas teórico-práticas.
- Código de conduta: (cf. Regulamento Disciplinar dos Estudantes do IPP)
    - Nenhum estudante ou grupo pode assumir pertença de trabalho realizado por outrem ou desenvolvido em conluio.
    - É expressamente proibido o uso de materiais, artefactos ou código de outrem sem a devida, e explícita indicação de origem.
    - Código de outras fontes deve ser claramente identificado no próprio código, indicando a fonte.
    - Casos de apropriação ilícita de materiais, artefactos e ou código sujeito a avaliação serão reportados à Presidência do ISEP.
    - A utilização de ferramentas com IA de assistência à codificação/desenho (e.g. chatGPT) deve ser mencionada.
- É obrigatório o uso da ferramenta de controle de versões Bitbucket.

3.1. Artigo Científico

No artigo científico deverão ser documentadas todas as fases da metodologia de trabalho seguida, preparação e exploração dos dados, análise e discussão dos resultados e conclusões (**máximo de 8 páginas** com o template do IEEE disponibilizado no moodle). Considere os seguintes aspetos:

- O artigo deverá ter uma secção inicial com o resumo de todas as técnicas estatísticas usadas na

realização da análise dos dados.

- O artigo deverá ter uma secção final com um resumo das principais conclusões retiradas da resolução das diferentes questões colocadas no enunciado.
- Cada problema resolvido deverá ter uma breve explicação da técnica estatística e uma conclusão baseada na interpretação e análise dos resultados obtidos. Caso considere relevante deverá incluir uma síntese destas conclusões na secção final de conclusões.

3.2. Avaliação

Na avaliação do trabalho serão considerados os seguintes aspetos e as ponderações especificadas na tabela 1:

a) Contextualização e objetivos (Sumário e Introdução)
b) Qualidade do código Phyton e respetiva documentação
c) A qualidade do processo de análise de dados seguido, a organização do código, a avaliação dos modelos criados e as conclusões alcançadas
d) Organização, qualidade da escrita, apresentação e clareza do relatório
e) A defesa e discussão
f) Participação individual de cada um dos elementos

Tabela 1 – Grelha de avaliação do Trabalho Prático 1

Sumário	15%
Questão 1	30%
Questão 2	20%
Questão 3	20%
Conclusão e referências	15%

Nota: A nota de cada um dos elementos do grupo será definida de acordo com a sua participação (em %). A equipa de avaliação de trabalhos práticos irá validar, no momento da defesa do trabalho (que poderá ser por videoconferência), a participação de cada um dos elementos do grupo na concretização dos objetivos do trabalho e do grupo. **Os elementos ausentes não terão classificação.**

4. Descrição do Trabalho

Na realização do Trabalho Prático 1 pretende-se que os alunos desenvolvam o processo de Análise Exploratória de Dados, Análise Inferencial, Correlação e Regressão [1-3].

Pretende-se no âmbito deste trabalho realizar a análise dados relativos a emissões de CO2 no período temporal do ano 1900 ao ano 2021, de diferentes regiões geográficas. Os dados necessários para a realização deste trabalho encontram-se no ficheiro **CO_data.csv** e o dicionário dos rótulos das colunas encontra-se no ficheiro **CO_data_dicionario.xlsx.** Use um grau de significância de 5% para todos os testes de hipótese efetuados.

4.1. Análise e exploração de dados

1. Construa um gráfico que permita visualizar as emissões totais de CO2 de Portugal no período 1900-2021. Encontre o ano em que foi emitido um valor máximo de CO2.
2. Construa um gráfico que permita comparar, no período 1900-2021, as emissões de CO2 de Portugal provenientes de: cimento, carvão, queima (flaring), gas, metano, óxido nitroso e do petróleo.
3. Construa um gráfico que permita comparar, no período 1900-2021, as emissões de CO2 per capita de Portugal com a Espanha.

4. Construa um gráfico que permita comparar as emissões de CO2 originadas pelo carvão dos Estados Unidos, China, Índia, União Europeia (a 27) e a Rússia no período 2000-2021.

5. Construa uma tabela que indique, para cada uma das regiões: Estados Unidos, China, Índia, União Europeia (a 27) e a Rússia, as médias das emissões de CO2 devidas a cimento, carvão, queima (flaring), gas, metano, óxido nitroso e do petróleo no período 2000-2021. (formate as entradas da tabela de forma a terem apenas 3 casas decimais).

4.2. Inferência Estatística

Suponha que não temos acesso a todos os dados relativos as anos 1900, 1901,...,2021, contudo temos acesso aos dados relativos a amostras aleatórias. Use um grau de significância de 5% para todos os testes de hipótese efetuados

1. Escolheu-se uma amostra aleatória (sampleyears1) da seguinte forma:

```
seed_value = 100
years = pd.Series([i for i in range(1900,2021)])
sampleyears1 = years.sample(n=30,replace=False)
```

Use os dados relativos aos anos da amostra sampleyears1 para testar se a média do produto interno bruto (gdp) de Portugal foi superior à média do produto interno bruto da Hungria no período 1900-2021.

2. Escolheram-se duas amostras aleatórias (**`sampleyears2`** e **`sampleyears3`**) da seguinte forma:

```
years = pd.Series([i for i in range(1900,2021)])
seed_value = 55
sampleyears2 = years.sample(n=12, replace=False)
seed_value = 85
sampleyears3 = years.sample(n=12, replace=False)
```

Use os dados relativos aos anos da amostra **`sampleyears2`** para Portugal e os dados relativos aos anos da amostra **`sampleyears3`** para a Hungria para testar se a média do produto interno bruto (gdp) de Portugal foi superior à média do produto interno bruto da Hungria no período 1900-2021.

3. Use os anos da amostra **`sampleyears2`** (ver questão anterior) para testar se há diferenças significativas nas emissões totais de CO2 entre as regiões: Estados Unidos, Rússia, China, India e União Europeia (a 27). **Nota:** Caso necessário efetue uma análise **post-hoc** adequada.

4.3. Correlação e Regressão

1. Use os dados da emissão de CO2 provenientes do carvão dos anos compreendidos entre 2000 e 2021 para construir uma tabela de correlação entre as regiões: África ('Africa'), Ásia ('Asia'), América do Sul ('South America'), América do Norte ('North America'), Europa ('Europe') e Oceania ('Oceania').
2. Considere as variáveis independentes:
    - X1 - Emissão de CO2 provenientes do carvão nos anos pares do século XXI na Alemanha
    - X2 - Emissão de CO2 proveniente do carvão nos anos pares do século XXI na Rússia
    - X3 - Emissão de CO2 proveniente do carvão nos anos pares do século XXI na França
    - X4 - Emissão de CO2 proveniente do carvão nos anos pares do século XXI em Portugal

e a variável dependente:

- Y - Emissão de CO2 provenientes do carvão nos anos pares do século XXI na Europa (’Europe’)

a) Encontre o modelo de regressão linear.

b) Verifique as condições sobre os resíduos.

c) Verifique se existe colinearidade (VIF).

d) Comente o modelo obtido tendo em conta todas as características relevantes para a qualidade do modelo.

e) Estime a emissão de CO2 proveniente do carvão na Europa no ano 2015 e compare com o valor real

4.4. Análise e Discussão de Resultados

Efetue uma síntese dos resultados e das conclusões, obtidos neste trabalho, que considera mais importantes, justificando sempre que necessário (conclusão).

5. Referências Bibliográficas

[1]. C. HEUMANN and SHALABH M. SCHOMAKER, Introduction to statistics and data analysis, Springer International Publishing, 2016.

[2]. DOUGLAS C. MONTGOMERY, Design and Analysis of Experiments, 8th edition. John Wiley & Sons, New York, 2013

[3]. WES MCKINNEY, PYTHON FOR DATA ANALYSIS: DATA WRANGLING WITH PANDAS, NUMPY, AND JUPYTER, 3RD EDITION, HTTPS://WESMCKINNEY.COM/BOOK/,O'REILLY MEDIA, 2022.