Trabalho Prático

Análise de Dados em Informática

Engenharia Informática - 3º ano 2º semestre
Ano Letivo 2020/2021

---

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 R
- Escrita de artigo científico com a Análise de Dados

2. Calendarização

Lançamento das propostas de trabalhos: até 20 de março de 2021
Entrega do trabalho: até **2 de maio de 2021** (23:55)
Defesa e discussão: em data a marcar pelo professor de TP

3. Normas

- O grupo deve ser o mesmo nas 2 iterações do trabalho prático.
- Deverá ser usado o R como ferramenta de suporte ao tratamento de dados.
- A **data final de ENTREGA** da 1ª iteração do trabalho é **2 de maio de 2021**, 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 conforme template disponibilizado no moodle,. 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 R para resolver o problema
- Deverá submeter todos os documentos num ficheiro compactado. 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.

3.1. Artigo Científico

No Artigo Científico deverão ser documentadas todas as fases da metodologia de trabalho seguida, contextualização do tema, exploração, preparação dos dados, análise e discussão dos resultados e conclusões.

3.2. Avaliação

Na avaliação do trabalho serão considerados os seguintes aspetos:

1. Contextualização e enquadramento teórico, motivação e objetivos (Introdução)
2. A qualidade do processo de análise de dados seguido, a organização do código, a avaliação dos modelos criados e as conclusões alcançadas
3. Organização, qualidade da escrita, apresentação e clareza do artigo científico
4. A defesa e discussão
5. Participação individual de cada um dos elementos

Introdução	**10%**

Análise Exploratória de Dados	**20%**
Inferência Estatística	**15%**
Correlação	**15%**
Regressão	**15%**
Conclusões	**15%**
Estrutura e organização do artigo	**10%**

Nota: A nota de cada um dos elementos do grupo será definida de acordo com a sua participação. A equipa de avaliação de trabalhos práticos irá validar, no momento da defesa do trabalho (que poderá ser via 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 deste trabalho pretende-se que os alunos desenvolvam o processo de Análise Exploratória de Dados[1], Análise Inferencial, Correlação e Regressão, do *dataset* facultado com dados reais sobre o COVID-19, retirados da base dados internacionais "*Our World in Data*" [2], dinamizada pela universidade *Johns Hopkins University (JHU)*.

O trabalho usa o ficheiro "owid-covid-data.xlsx", disponível no moodle, retirado do link "https://ourworldindata.org/coronavirus-source-data". O ficheiro contém dados relativos ao período 2020-01-01 até 2021-02-27.

Importante: é necessário filtrar os dados e retirar os NA.

4.1. Análise de dados

a) Gráfico que mostra o número total de infetados ao longo do período de tempo (indicado no ficheiro de dados), por continente.

b) Gráfico do total de infetados por milhão de habitantes, ao longo do período de tempo, por continente.

c) Um *boxplot* do número de mortos diários por milhão de habitantes para cada um dos seguintes países: Portugal, Espanha, Itália e Reino Unido. Remova os *outliers* usando o critério: **x** é *outlier* sse **x** $\notin$ [Q1-1.5*IQR , Q3+1.5*IQR].

d) Um gráfico de barras com o número total de mortos, por milhão de habitantes, e o nº de testes diários por milhar de habitantes, para os países: Albânia, Dinamarca, Alemanha e Rússia.

e) Indique qual o país europeu que teve o maior número de infetados, por milhão de habitantes, num só dia.

f) Indique em que dia, e em que pais, se registou a maior taxa de transmissibilidade do vírus.

g) Efetue um *boxplot* do nº de mortos diários por milhão de habitantes, em cada continente. Remova os outliers usando o critério: **x** é *outlier* sse **x**$\notin$ [Q1-1.5*IQR , Q3+1.5*IQR].

4.2. Inferência estatística

Para a geração de amostras pseudoaleatórias, pertencentes ao período 2020-04-01 até 2021-02-27, considere o uso da função *set.seed()*:

a) Considerando apenas os dados relativos, a 30 dias, da amostra pseudoaleatória (usando o valor 118 no parâmetro da função set.seed()) que obteve, verifique se a média da taxa transmissibilidade no Reino Unido é superior à média da taxa de transmissibilidade em Portugal.

b) Considerando apenas os dados relativos, a 15 dias, da amostra pseudoaleatória (usando o valor 115 no parâmetro da função set.seed()) que obteve, verifique se há diferenças significativas entre o nº de

mortes diárias, por milhão de habitantes, em Espanha, França, Portugal e Itália. No caso de haver, efetue uma análise *post-hoc*.

c) Para cada Continente gere uma amostra pseudoaleatória, de 30 dias. Para a África use a *seed* 100, para a Ásia use a *seed* 101, para a Europa use a *seed* 102, para a América do Norte use a *seed* 103, e para a América do Sul use a *seed* 104. Verifique se existe diferença significativa entre os números médios diários de mortes, por milhão de habitantes, entre os continentes. No caso de haver, efetue uma análise *post-hoc*.

4.3. Correlação

Averigue se existe correlação, em 2021, entre:

a) o valor máximo da taxa diária de transmissibilidade e a densidade populacional de todos os países da Europa com mais de 10 milhões de habitantes.

b) o total de mortos por milhão de habitantes e a percentagem da população com 65 anos ou mais em todos os países da Europa com mais de 10 milhões de habitantes.

Comente devidamente todos os pressupostos que assumir e os resultados obtidos.

4.4. Regressão

Considere o "Índice de rigor" médio mensal (**Ir**, Stringency Index) em Portugal, a variável dependente e as variáveis independentes: média de mortes diárias por milhão de habitantes $\mathbf{D_m}$, média de casos diários por milhão de habitantes $\mathbf{C_m}$ e a média mensal da taxa de transmissibilidade $\mathbf{R_m}$. Considere os dados no período 2020-04-01 até 2021-02-27.

a) Construa o modelo de regressão linear múltipla.

b) Verifique se as condições de: Homocedasticidade, Autocorrelação nula e de Multicolinearidade são satisfeitas.

c) Estime o valor de **Ir** para os valores $\mathbf{D_m} = 10$, $\mathbf{C_m} = 460$ e $\mathbf{R_m} = 1.1$

5. Referências Bibliográficas

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

[2]. Coronavirus Source Data - Our World in Data, https://ourworldindata.org/coronavirus-source-data