Prova pacial – Data wrangling – 20 / 10 /2020

**Nome\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Ao entregar este trabalho final no e-class Você estará declarando:**

**"Estou ciente das razões pelas quais estamos fazendo este trabalho de forma remota. Assim, declaro que utilizei somente formas éticas de resolver as questões e não consultei ninguém, de nenhuma forma, para respondê-las ”**

Escrevi para que Você leia, por favor!!!

O não cumprimento das instruções será penalizado

**Instruções**:

* Este teste tem **6 páginas**, incluindo esta.
* Consulta a notas de aula, apontamentos, net, scripts etc. é permitida.
* Trabalho é individual. A consulta a colegas ou outras pessoas, qualquer que seja o meio, é proibida.
* Todos os cálculos e gráficos deverão ser feitos com o software R
* Responda as perguntas digitando / colando as respostas neste documento, nos espaços indicados. Respostas fora desses espaços não serão consideradas.
* Por favor, só responda o que for pedido. Não perca tempo tentando explicar o que fez ou por que fez (mesmo porque eu não lerei).
* **Na correção, dependendo da questão, não existirá “meio certo”. (se o comando está errado, a resposta numérica não pode dar certo! Se o comando estiver correto, a resposta numérica não poderia dar errada!!)**
* Se Você copiar algum output do R, formate-o como *Lucida Console*. Cuidado: se não formatar apresentação poderá ficar feia e sua nota será penalizada. É mais fácil copiar os comandos de programação da caixinha de script do que de output
* Os gráficos tem que caber nos espaços indicados
* Ao terminar, **não esqueça de colocar seu nome**. Verifique cuidadosamente se não esqueceu nenhuma resposta nos espaços indicados.
* **Salve como .pdf** e faça upload no ECLASS na caixa de **entrega de atividades**, na pasta correspondente. Verifique se o arquivo salvo ficou em “boas condições” (ou seja se não foi desformatado. Pode acontecer) **Provas entregues no formato .doc não serão aceitos** .
* Veja que a data de entrega é **até as 10h28min do dia 27/10/2020**
* Não envie as respostas por e mail ou WhatsApp. Não serão aceitas

**Questão 1:** Considere a planilha **WH\_2018** no arquivo **DW PARCIAL 2020 10 20** em Excel postado no eclass.

* Não altere esta planilha a menos que seja solicitado.
* Não elimine observações, a menos que seja solicitado
* Para simplificar batize esse arquivo como **ww= WH\_2018**

**Questão 1.A:** Elabore um boxplot com o arquivo original WH\_18 para analisar a variável LifeLadder. Coloque o título no gráfico e faça o grid em cor vermelha (red). Dê os valores da mediana, 1º e 3º quartil do boxplot.

Cole o(s) comando(s) do R e o boxplot nos espaços abaixo. O gráfico tem que caber na caixa.

Comando(s) do gráfico

Boxplot

Comando de cálculo das mediana, quartis (3 casas decimais) e respostas

**Questão 1.B** : Discretize a variável LifeLadder em três categorias de mesma frequência, gerando a variável klife. Depois obtenha a tabela mostrando os limites dessas categorias e a frequência relativa de cada uma. As frequências com apenas 3 casa decimais

Cole os Comandos

Cole a Tabela. **Utilize 3 casa decimais apenas**

**Questão 1.C** : calcule a média das variáveis LifeLadder e LogGDPpercapita. Dê as respostas com duas casas decimais

Média de LifeLadder: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (2 casas decimais)

Média de LogGDPpercapita: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (2 casas decimais)

**Questão 1.D**: Considere a variável Positiveaffect que apresenta um missing value. Estime o valor do missing value utilizando a opção de kNN do arquivo VIM, considerando **10** vizinhos, e com as opções addRandom = F e weightDist = F. Identifique com o comando which em que linha está o missing value.

Comandos do R

Valor imputado (3 casas decimais)

**Questão 1.E**: Discretize a variável LifeLadder em três categorias: de 0,0 até 4,0, de 4,0 a 6,0 e de 6,0 até 10,0. Denomine esta nova variável como **klad**. Construa uma tabela mostrando as frequências absolutas dentro de cada categoria. Calcule a média de LifeLadder em cada uma dessas categorias utilizando a função tapply do R.

|  |  |  |
| --- | --- | --- |
| Limites das classes | Frequência absoluta | média de LifeLadder na classe  (arredonde com 3 casas decimais) |
|  |  |  |
|  |  |  |
|  |  |  |

**Continua . 2ª questão começa na página seguinte**

**Questão** **2**: Considere a planilha **WH\_ALLYEARS** no arquivo **DW PARCIAL 2020 10 20** em Excel postado no eclass.

* Para simplificar batize esse arquivo como **all= WH\_ALLYEARS**

**Questão 2.A)** Utilizando as funções adequadas do dplyr e aplicando os operadores pipe (%>%) crie o arquivo **dd5** que contenha apenas os dados do ano 2005 (mantendo todas as colunas originais). Após isto, calcule a média de logGDPpercapita em **dd5**

Comandos para gerar dd5(uma linha)

Comandos para cálculo da média (uma linha)

Calcule dim(dd5)=

Média de logGDPpercapita (3 casas decimais)

**Questão 2.B)** Utilizando as funções adequadas do dplyr e aplicando os operadores pipe gere o arquivo **ppp** que contenha apenas os dados do ano 2005 e os países que nesse ano tiveram um LifeLadder superior 6,0 (**ppp** mantendo todas as colunas originais).

Após isto calcule a média de logGDPpercapita para o arquivo DD2. (3 casas decimais)

Comandos (geração do ppp e cálculo da média )

Média de logGDPpercapita (3 casas decimais)

**Questão 2.C:** Utilizando as funções adequadas do dplyr, selecione apenas os anos 2005, 2006 e 2007, (ou seja Year<=2007) agrupe os dados por ano e calcule as médias de LifeLadder para cada ano. Utilize **apenas uma linha** de comando (copie direto do script) para responder toda esta questão. Comece por all%>%....

Comando (copie direto do script)

all %>% etc.....

|  |  |
| --- | --- |
| Ano | Média anual de LifeLadder (2 casas decimais) |
| 2005 |  |
| 2006 |  |
| 2007 |  |

**Questão 2.D:** utilizando as funções de dplyr como o operador pipe, gere o dataset **mm** contendo apenas o ano 2007, e contendo apenas as variáveis LifeLadder, logGDOpercapita e Socialsupport. A seguir, nesse mesmo arquivo, calcule a variável **RAIZ**= raiz quadrada de Socialsupport . **O comando que Você deve utilizar para responder (até aqui) esta questão deve ter apenas uma linha no Script do R** (copie o comando da janela de Script. Pode ser que ao passar para a caixa de resposta baixo talvez “quebre” em duas linhas para caber. Isto não será problema). Finalmente, com outro comando, imprima as 8 primeiras linhas de **mm**.

8 primeiras linhas (copie do output, formate como Lucida Console)