FIAP

Descriptive & Predictive Analytics – Turma 1DTS

Trabalho 1

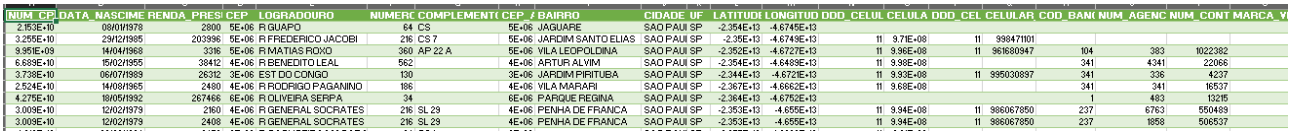
Habilidades desenvolvidas: Analisar tabelas de frequências e gráficos. Análise exploratória dos dados. Detecção de outliers.

1) Classifique o tipo de variável

Utilize a base de dados do cadastro para classificar o tipo de variável.

Variável	Tipo da variável (natureza)
NUM_CPF	ID
CHAVE_CONTRATO	ID
DATA_NASCIMENTO	Qualitativa ordinal
RENDA_PRESUMIDA	Quantitativa contínua
CEP	Qualitativo nominal
LOGRADOURO	Qualitativo nominal
NUMERO	Qualitativo ordinal
COMPLEMENTO	Qualitativo nominal
CEP_A	Qualitativo nominal
BAIRRO	Qualitativo nominal
CIDADE	Qualitativo nominal
UF	Qualitativo nominal
LATITUDE	Qualitativo nominal
LONGITUDE	Qualitativo nominal
DDD_CELULAR	Qualitativo nominal
CELULAR	Qualitativo nominal
DDD_CELULAR_2	Qualitativo nominal
CELULAR_2	Qualitativo nominal
COD_BANCO	Qualitativo nominal
NUM_AGENCIA	Qualitativa nominal
NUM_CONTA	ID
MARCA_VEICULO	Qualitativa nominal
MODELO_VEICULO	Qualitativa nominal
PLACA_VEICULO	Qualitativa nominal
CNPJ_CREDOR	Qualitativa nominal
VALOR_DIVIDA	Quantitativa contínua
STATUS_CONSENTIMENTO	Qualitativa nominal
DATA_INCLUSAO1	Qualitativa ordinal
IDADE	Quantitativa contínua

Trecho do arquivo Cadastro_PF.csv



NUM_CP	DATA_NASCIME	RENDA_PRES	CEP	LOGRADOURO	NUMERC	COMPLEMENTO	CEP_2	BAIRRO	CIDADE	UF	LATITUDE	LONGITUDE	DDD_CELULA	CELULA	DDD_CELULAR	CELULAR	COD_BANCO	NUM_AGENC	NUM_CONTA	MARCA_VE
2.153E+10	08/01/1978	2800	5E+06	R GUAPO	64	CS	5E+06	JAGUARE	SAO PAULO	SP	-2.354E+13	-4.675E+13								
3.255E+10	29/12/1985	203996	5E+06	R FREDERICO JACOBI	216	CS 7	5E+06	JARDIM SANTO ELIAS	SAO PAULO	SP	-2.35E+13	-4.674E+13	11	9.71E+08	11	998471101				
9.951E+09	14/04/1968	3316	5E+06	R MATIAS ROXO	360	AP 22 A	5E+06	VILA LEOPOLDINA	SAO PAULO	SP	-2.352E+13	-4.672E+13	11	9.96E+08	11	961680947	104	383	1022382	
6.689E+10	15/02/1955	38412	4E+06	R BENEDITO LEAL	562		4E+06	ARTUR ALVIM	SAO PAULO	SP	-2.354E+13	-4.648E+13	11	9.38E+08			341	4341	22066	
3.738E+10	06/07/1989	26312	3E+06	EST DO CONGO	130		3E+06	JARDIM PIRITUBA	SAO PAULO	SP	-2.344E+13	-4.672E+13	11	9.33E+08	11	995030897	341	336	4237	
2.524E+10	14/08/1965	2480	4E+06	R RODRIGO PAGANINO	186		4E+06	VILA MARARI	SAO PAULO	SP	-2.367E+13	-4.666E+13	11	9.68E+08			341	341	16537	
4.275E+10	18/05/1992	267466	6E+06	R OLIVEIRA SERPA	34		6E+06	PARQUE REGINA	SAO PAULO	SP	-2.364E+13	-4.675E+13					1	483	13215	
3.009E+10	12/02/1979	2160	4E+06	R GENERAL SOCRATES	216	SL 29	4E+06	PENHA DE FRANCA	SAO PAULO	SP	-2.353E+13	-4.655E+13	11	9.94E+08	11	986067850	237	6763	550489	
3.009E+10	12/02/1979	2408	4E+06	R GENERAL SOCRATES	216	SL 29	4E+06	PENHA DE FRANCA	SAO PAULO	SP	-2.353E+13	-4.655E+13	11	9.94E+08	11	986067850	237	1858	506537	

2) Relacione a Estatística com a Definição do conceito estatístico de cada medida resumo:

	Estatísticas		Definição
(a)	Média	(G)	Medida de dispersão que caracteriza o pico ou "achatamento" da curva da função de distribuição de probabilidade
(b)	Erro padrão	(K)	É o maior valor de um conjunto de dados.
(c)	Mediana	(J)	É o menor valor de um conjunto de dados.
(d)	Moda	(C)	Valor numérico que separa a metade superior de uma amostra de dados, uma população ou uma distribuição de probabilidade, a partir da metade inferior.
(e)	Desvio padrão	(D)	Valor que detém o maior número de observações, ou seja, o valor ou valores mais frequentes, ou ainda "o valor que ocorre com maior frequência num conjunto de dados, isto é, o valor mais comum
(f)	Variância da amostra	(B)	Medida de variabilidade da média amostral.
(g)	Curtose	(I)	Medida de dispersão. É a diferença entre o máximo e o mínimo.
(h)	Assimetria	(A)	Valor que aponta para onde mais se concentram os dados de uma distribuição. Pode ser considerada o ponto de equilíbrio das frequências, num histograma.
(i)	Amplitude	(F)	Medida que se obtém somando os quadrados dos desvios dos dados relativamente à média, e dividindo pelo número de dados menos um. Representa-se por s2.
(j)	Mínimo	(E)	É a medida mais comum da dispersão estatística. Ele mostra a quanto de variação ou dispersão em relação à média.
(k)	Máximo	(L)	Conjunto constituído pela reunião de diversos subconjuntos; total, conjunto, somatório.
(l)	Soma	(H)	Grau de desvio de afastamento da simetria de uma distribuição, pode ser positiva para distribuições a direita e negativa para a esquerda. Para distribuições simetricas seu valor é zero.
(m)	Contagem	(M)	Quantidade de registros.

3) Critique a tabela abaixo:

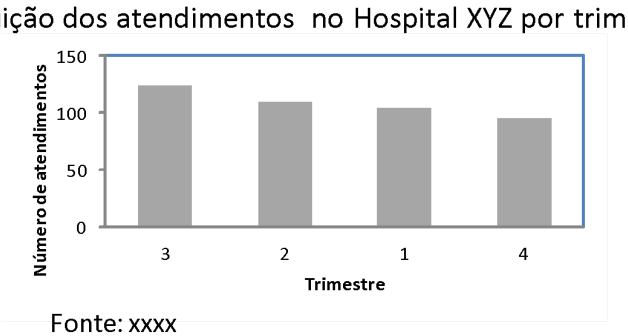
Salários pagos na empresa XYZ (2o. semestre de 2009)

Meses/ano	No. de Empregados	Salários Pagos
Jul/09	57	30.127
Ago/09	61	32.893
Set/09	58	31.041
Out/09	55	29.412
Nov/09	66	35.523
Dez/09	75	40.080

Total	372	199.076

R: FALTA DA UNIDADE NA COLUNA “SALÁRIOS PAGOS”. APÓS O ENTENDIMENTO DO SIGNIFICADO DA COLUNA “SALÁRIOS PAGOS”, É POSSÍVEL ADICIONAR MEDIDAS DE RESUMO, COMO MÉDIA, MEDIANA E DESVIO PADRÃO PARA REALIZAR ESTUDOS DA VARIÁVEL.

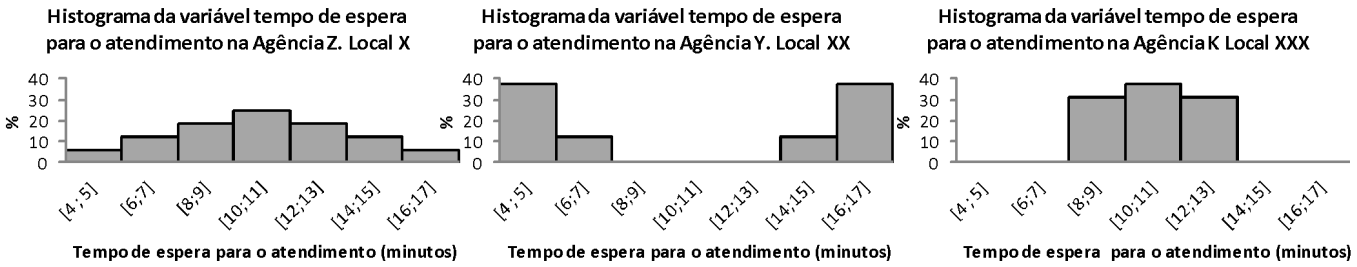
4) UM GRÁFICO PODE LEVAR A CONCLUSÕES ERRADAS? EXPLIQUE POR QUE O GRÁFICO ABAIXO LEVA A CONCLUSÕES ERRADAS.



Fonte: xxxx

R: SIM, A ALTERAÇÃO DA ORDEM DOS SEMESTRES, DA IMPRESSÃO QUE QUE O NÚMERO DE ATENDIMENTOS VEM DIMINUINDO, MAS SE ORDENARMOS A ORDEM DO TRIMESTRE, A ANÁLISE SERÁ QUE NO ATÉ O 3 TRIMESTRE TEVE UM AUMENTO DO NÚMERO DE ATENDIMENTOS MÉDICOS E NO QUARTO TREMESTRE HOUVE UMA QUEDA EM RELAÇÃO AO SEMESTRE ANTERIOR.

5) SEM CALCULAR, OBSERVE OS TRÊS GRÁFICOS E RESPONDA?



A) QUAL É O CONJUNTO DE DADOS COM MAIOR DESVIO PADRÃO? QUAL TEM O MENOR DESVIO PADRÃO? EXPLIQUE SEU RACIOCÍNIO.

R: MAIOR DESVIO PADRÃO: HISTOGRAMA 2 - AGÊNCIA Y, pois a maior parte dos dados estão mais distantes da média. O MENOR DESVIO PADRÃO É O HISTOGRAMA 3 - AGENCIAZ, POIS OS DADOS ESTÃO MAIS PRÓXIMOS DA MÉDIA.

B) NO QUE OS CONJUNTOS DE DADOS SÃO IGUAIS? NO QUE ELES DIFEREM?

R: Os conjuntos de dados são iguais nas médias e se diferem no desvio padrão.

C) EM QUAL AGÊNCIA VOCÊ IRIA? JUSTIFIQUE?

R: Na agência Y, pois a chances de ocorrer um atendimento entre 4 e 5 minutos é maior do que nas demais agências.

6) BOX PLOT

A variável RENDA_PRESUMIDA tem outlier ou ponto extremo?

Use o Box Plot para detectar presença de outlier e ponto extremo. Apresente os cálculos para identificar os limites que definem os outliers e pontos extremos.

```
> summary(RENDA_PRESUMIDA)
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
      0    2068    2432   10388    2924   1890000
```

Mínimo = 0

Máximo = 1890000

Quartil 1 (Q1) = 2068

Quartil 2 (Q2) = mediana = 2.432

Quartil 3 (Q3) = 2924

Interquartil (IQ) = Q3 – Q1 = 856

Limite 1 = Q3 + 3*IQ = 5.492

Limite 2 = Q3 + 1;5*IQ = 4.208

Limite 3 = Q1 – 1.5*IQ = 784

Limite 4 = Q1 – 3*IQ = -500 = Ø

R: Podemos perceber um número alto de outliers já observando as medidas limite 1, 2, 3 e 4.Acima do limite 1 seriam os valores do pontos extremo, entre limite 1 e 2 e entre o limite 3 e 4 seriam os outliers.

Entrega do exercício no formato word.

Data de entrega: 12/05/2022

Regina Bernal

FIAP

27/04/2022