**Conceitos Estatísticos para IA – Turma 20IA**

**Trabalho 3**

**Habilidades desenvolvidas: Análise exploratória dos dados. Análise de correlação e associação. Modelo preditivo usando a regressão linear simples**

1. Faça a previsão das vendas (R$) mensal no período de 12 meses da empresa XYZ a partir dos dados disponíveis de Vendas (R$) e Budget Advertising (R$) da empresa (arquivo: Vendas\_2016a2018.csv). Use o modelo de regressão linear simples:

|  |  |  |  |
| --- | --- | --- | --- |
| Data | ano | Vendas | Budget\_Advertising |
| jan/16 | 2016 | 1160081 | 72800 |
| fev/16 | 2016 | 1622540 | 123392 |
| mar/16 | 2016 | 1597260 | 135761 |
| abr/16 | 2016 | 1640675 | 148064 |
| mai/16 | 2016 | 1511270 | 159746 |
| jun/16 | 2016 | 1634073 | 183353 |
| jul/16 | 2016 | 1856971 | 190722 |
| ago/16 | 2016 | 1585566 | 197802 |
| set/16 | 2016 | 2041672 | 248891 |
| out/16 | 2016 | 1933557 | 256353 |
| nov/16 | 2016 | 2076910 | 298805 |
| dez/16 | 2016 | 1740202 | 268925 |
| jan/17 | 2017 | 1657258 | 104000 |
| fev/17 | 2017 | 2317914 | 176274 |
| mar/17 | 2017 | 2281800 | 193944 |
| abr/17 | 2017 | 2343821 | 211520 |
| mai/17 | 2017 | 2158957 | 228209 |
| jun/17 | 2017 | 2334390 | 261933 |
| jul/17 | 2017 | 2652815 | 272460 |
| ago/17 | 2017 | 2265094 | 282574 |
| set/17 | 2017 | 2916674 | 355559 |
| out/17 | 2017 | 2762225 | 366219 |
| nov/17 | 2017 | 2967014 | 426865 |
| dez/17 | 2017 | 2486002 | 384179 |
| jan/18 | 2018 | 2071573 | 130000 |
| fev/18 | 2018 | 2897392 | 220343 |
| mar/18 | 2018 | 2852250 | 242431 |
| abr/18 | 2018 | 2929776 | 264401 |
| mai/18 | 2018 | 2698696 | 285262 |
| jun/18 | 2018 | 2917988 | 327417 |
| jul/18 | 2018 | 3316019 | 340575 |
| ago/18 | 2018 | 2831367 | 353218 |
| set/18 | 2018 | 3645842 | 444449 |
| out/18 | 2018 | 3452781 | 457774 |
| nov/18 | 2018 | 3708768 | 533582 |
| dez/18 | 2018 | 3107503 | 480224 |
| jan/19 | 2019 |  | 91000 |
| fev/19 | 2019 |  | 154240 |
| mar/19 | 2019 |  | 169702 |
| abr/19 | 2019 |  | 185081 |
| mai/19 | 2019 |  | 199683 |
| jun/19 | 2019 |  | 229192 |
| jul/19 | 2019 |  | 238403 |
| ago/19 | 2019 |  | 247253 |
| set/19 | 2019 |  | 311114 |
| out/19 | 2019 |  | 320442 |
| nov/19 | 2019 |  | 373507 |
| dez/19 | 2019 |  | 336157 |

1. Definição do problema:
2. Objetivo do estudo:
3. Classifique os tipos de variáveis:

|  |  |  |
| --- | --- | --- |
| Variável | Tipo da variável | Regra de data mining |
| Data |  |  |
| Ano |  |  |
| Vendas |  |  |
| Budget Advertising |  |  |

1. Qual a variável resposta do modelo? Qual a variável preditora?
2. Apresente o gráfico de dispersão entre Vendas e Budget. Interprete o resultado.
3. Qual a correlação entre Vendas e Budget? Interprete o resultado.
4. Apresente a saída da regressão linear simples. Interprete os resultados.
5. Qual a acurácia do modelo ajustado?
6. Apresente a análise de resíduos. Interprete os resultados.
7. Qual a previsão de Vendas para 2019? Utilize o arquivo Budget\_2019.csv.
8. Use os dados *Bike\_Sharing.csv* para construir as análises descritivas, correlação, associação e modelo preditivo para previsão do número de bikes alugadas por mês.

Descrição:

Os sistemas de compartilhamento de bicicletas são uma nova geração de aluguel de bicicletas tradicional, onde todo o processo de associação, locação e devolução tornou-se automático. Através destes sistemas, o usuário pode facilmente alugar uma bicicleta a partir de uma determinada posição e retornar em outra posição. Atualmente, existem cerca de 500 programas de compartilhamento de bicicletas em todo o mundo, compostos por mais de 500 mil bicicletas. Hoje, existe um grande interesse nesses sistemas devido ao seu importante papel no trânsito, questões ambientais e de saúde.

Fonte de dados: <https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>

Atividades:

1. Classifique o tipo de variável

|  |  |  |  |
| --- | --- | --- | --- |
| **Variável** | **Descrição** | **Tipo de Variável** | **Regra de data mining** |
| **instant** | record index | ID |  |
| **dteday** | date | ID |  |
| **season** | season (~~1:springer, 2:summer, 3:fall, 4:winter~~)  1:winter, 2:spring, 3:summer, 4:fall | Qualitativa Ordinal | preditora |
| **yr** | year (0: 2011, 1:2012) | Qualitativa Ordinal | preditora |
| **mnth** | month ( 1 to 12) | Qualitativa Ordinal | preditora |
| **hr** | hour (0 to 23) |  | preditora |
| **holiday** | weather day is holiday or not | Qualitativa Ordinal | preditora |
| **weekday** | day of the week | Qualitativa Ordinal | preditora |
| **workingday** | if day is neither weekend nor holiday is 1, otherwise is 0. | Qualitativa nominal | preditora |
| **weathersit** | 1: Clear, Few clouds, Partly cloudy, Partly cloudy; 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist;3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds;4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog | Qualitativa Ordinal | preditora |
| **temp** | Normalized temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-8, t\_max=+39 (only in hourly scale) | Quantitativa | preditora |
| **atemp** | Normalized feeling temperature in Celsius. The values are derived via (t-t\_min)/(t\_max-t\_min), t\_min=-16, t\_max=+50 (only in hourly scale) | Quantitativa | preditora |
| **hum** | Normalized humidity. The values are divided to 100 (max) | Quantitativa | preditora |
| **windspeed** | Normalized wind speed. The values are divided to 67 (max) | Quantitativa | preditora |
| **casual** | count of casual users | Quantitativa | Target |
| **registered** | count of registered users | Quantitativa | Target |
| **cnt** | count of total rental bikes including both casual and registered | Quantitativa | Target |

1. Definição do problema:
2. Objetivo do estudo:

Obs: escolha uma variável resposta: cnt ou casual ou registered.

1. Apresente as tabelas descritivas e gráficos das variáveis. Interprete os resultados.
2. Quais variáveis foram normalizadas? Por quê? Apresente a fórmula utilizada.

Para as variáveis preditoras estarem em escalas semelhantes. A grandeza da variável influência no modelo estatístico.

1. Apresente a análise de associação e correlação de Pearson. Quais variáveis têm correlação com a variável resposta?
2. Criar o script R para o cálculo da variável Zscore da ingestão de sal. Utilize a base de dados “sal\_lab1.xlsx”.
3. Faça o histograma da variável original e Zscore.

library(readxl)

library(dplyr)

data <- read\_excel("fiap/desafios-fiap/Estatísticas/trab3/sal\_lab1.xlsx")

head(data)

mean(data$Ingestao\_Sal)

sd(data$Ingestao\_Sal)

data\_with\_zscore <- data %>%

mutate(zscore= (Ingestao\_Sal - mean(Ingestao\_Sal)/sd(Ingestao\_Sal)))

data\_with\_zscore

par(mfrow=c(1,2))

par(mar=c(10,4,8,2))

hist(data$Ingestao\_Sal)

hist(data\_with\_zscore$zscore)

1. Qual a diferença entre as duas variáveis?

Data de entrega: 10/09/2021

Regina Bernal

26/08/2021