PROJECT: Hospital Cost Report Public Use File

Financial Analysis with SQL Language and Linear Regression in R Language

1. Working Directory

Configurando o diretório de trabalho

setwd("C:/Users/Utilizador/repos/Formacao_cientista_de_dados/big_data_analytics_R_microsoft_azure_machine_lear
getwd()

2. Imports library(dplyr) library(sqldf)

3. Data Loading dados <- read.csv('dataset.csv')

4. Data Description

Descreve, compreende, organiza e resumi os dados

4.1 Data Viewer

Visualiza os dados

View(dados)

4.2 Data Dimension

Dimensões

dim(dados)

4.3 Data Type

Variáveis e tipos de dados

str(dados)

4.4 Data Summary

Sumários das variáveis numéricas

summary(dados)

5. Data Cleaning

Limpeza dos Dados

5.1 Check NA

Verificando ocorrência de valores NA

colSums(is.na(dados))

5.2 Fill Na

Apenas 1 registro com valor NA. Vamos removê-lo.

dados <- na.omit(dados)

Verificando ocorrência de valores NA

colSums(is.na(dados))

Dimensões

dim(dados)

Tipos de dados

str(dados)

6. Descriptive Statistivas 6.1 Numerical Attributes

Extraindo as variáveis numéricas

numeric_variable_list <- sapply(dados, is.numeric) numerical_data <- dados[numeric_variable_list]

Matriz de Correlação das variavéis numéricas

cor(numerical_data)

Correlation Plot

Plot com todas as variaveis numéricas

pairs(numerical_data)

7. Business Questions

7.1 Análise Exploratória com Linguagem SQL

Função usada para fazer consulta sql em um dataframe

?sqldf

Nomes das colunas

names(dados)

1- Quantas raças estão representadas no dataset?

num_racas <- sqldf('SELECT RACE, COUNT(RACE) as Num_Races FROM dados GROUP BY RACE') num_racas

2- Qual a idade média dos pacientes?

idade_media <- sqldf('SELECT AVG(AGE) as Idade_Media FROM dados') idade_media

3- Qual a moda da idade dos pacientes?

idade_moda <- sqldf(" SELECT AGE as Idade FROM (SELECT AGE, COUNT(AGE) AS count_age FROM dados GROUP BY AGE ORDER BY count_age DESC) LIMIT 1") idade_moda

4- Qual a variância da coluna idade?

idade_variancia <- sqldf("SELECT SUM((AGE - (SELECT AVG(AGE) FROM dados)) * (AGE - (SELECT AVG(AGE) FROM dados))) / (COUNT(AGE) - 1) AS variancia FROM dados") idade_variancia

5- Qual o gasto total com internações hospitalares por idade?

gasto_total_idade <- sqldf(' SELECT AGE as Idade, SUM(TOTCHG) as Gasto_Total FROM dados GROUP BY AGE') gasto_total_idade

6- Qual idade gera o maior gasto total com internações hospitalares?

Verificando a classe do objeto

class(gasto_total_idade)

Reorganizando os dados

Ordenando

Devolvendo apenas a primeira linha

arrange(gasto_total_idade, desc(Gasto_Total))[1,]

7- Qual o gasto total com internações hospitalares por gênero?

1 feminino

0 masculino

gasto_total_genero <- sqldf(' SELECT FEMALE as Genero, SUM(TOTCHG) as Gasto_Total FROM dados GROUP BY FEMALE') gasto_total_genero

8- Qual a média de gasto com internações hospitalares por raça do paciente?

gasto_medio_race <- sqldf(' SELECT RACE, AVG(TOTCHG) as Gasto_Medio FROM dados GROUP BY RACE') gasto_medio_race

9- Para pacientes acima de 10 anos, qual a média de gasto total com internações hospitalares?

gasto_medio_idade_acima_10anos <- sqldf(' SELECT AGE as Idade, AVG(TOTCHG) as Gasto_Medio FROM dados WHERE AGE > 10 GROUP BY AGE') gasto_medio_idade_acima_10anos

10- Considerando o item anterior, qual idade tem média de gastos superior a 3000?

where: filtragem de registros sobre colunas existentes

having: filtragem de registros sobre operações (sum, avg)

gasto_medio_idade_acima_10anos_acima3k <- sqldf(' SELECT AGE as Idade, AVG(TOTCHG) as Gasto_Medio FROM dados WHERE AGE > 10 GROUP BY AGE HAVING AVG(TOTCHG) > 3000') gasto_medio_idade_acima_10anos_acima3k

7.2 Análise (Estatística) de Regressão com Linguagem R

Pergunta 1:

Qual a distribuição da idade dos pacientes que frequentam o hospital?

Resposta: Crianças entre 0 e 1 ano são as que mais frequentam o hospital.

hist(dados$AGE)

hist(dados$AGE, main = "Histograma da Distribuição da Idade dos Pacientes que Frequentam o Hospital", xlab = "Idade", border = "black", col = c("light green", "dark green"), xlim = c(0,20), ylim = c(0,350))

Se fazemos o summary com variável do tipo numérico, o resultado é um resumo estatatístico.

summary(dados$AGE)

Convertemos a variável para o tipo fator e então obtemos o sumário que precisamos.

resumo_idade <- summary(as.factor(dados$AGE)) resumo_idade

Pergunta 2:

Qual faixa etária tem o maior gasto total no hospital?

Resposta: Crianças entre 0 e 1 ano são as que geram maior gasto no hospital.

?aggregate gasto_total_baseado_idade = aggregate(TOTCHG ~ AGE, FUN = sum, data = dados)
View(gasto_total_baseado_idade)

Buscando o maior valor

which.max(tapply(gasto_total_baseado_idade$TOTCHG, gasto_total_baseado_idade$TOTCHG, FUN = sum))

Visualizando o resultado

barplot(tapply(gasto_total_baseado_idade$TOTCHG, gasto_total_baseado_idade$AGE, FUN = sum))

Pergunta 3:

Qual grupo baseado em diagnóstico (Aprdrg) tem o maior gasto total no hospital?

Resposta: O grupo 640 tem o maior gasto total.

gasto_total_baseado_diag = aggregate(TOTCHG ~ APRDRG, FUN = sum, data = dados)
View(gasto_total_baseado_diag)

Filtrando o dataframe:

linha: valor maximo

colunas: todas as colunas

gasto_total_baseado_diag[which.max(gasto_total_baseado_diag$TOTCHG),]
* ANOVA

Pergunta 4:

A raça do paciente tem relação com o total gasto em internações no hospital?

Resposta: O valor-p é maior que 0.05. Falhamos em rejeitar a H0.

A raça do paciente não influencia no gasto total com internação no hospital.

Usaremos um Teste ANOVA.

Variável dependente no Teste ANOVA: TOTCHG (antes do ~)

Variável independente no Teste ANOVA: Race (depois do ~)

H0: Não há efeito de RACE em TOTCHG.

H1: Há efeito de RACE em TOTCHG.

Resumo da variavél Race(raça): tipo inteiro

```
summary(dados$RACE)
```

Convertendo a variavel Raca para factor e resumindo: tipo factor

```
summary(as.factor(dados$RACE))
```

Modelo anova

```
modelo_anova_1 <- aov(TOTCHG ~ RACE, data = dados)
```

Resumo do modelo

```
summary(modelo_anova_1)
```

Pergunta 5:

A combinação de idade e gênero dos pacientes influencia no gasto total em internações no hospital?

Resposta: Em ambos os casos o valor-p é menor que 0.05. Rejeitamos a hipótese nula.

Há um efeito significativo da idade e do gênero nos custos hospitalares.

Usaremos um Teste ANOVA.

Variável dependente no Teste ANOVA: TOTCHG (antes do ~)

Variáveis independentes no Teste ANOVA: AGE, FEMALE (depois do ~)

H0: Não há efeito de AGE e FEMALE em TOTCHG.

H1: Há efeito de AGE e FEMALE em TOTCHG.

Modelo anova

```
modelo_anova_2 <- aov(TOTCHG ~ AGE + FEMALE, data = dados)
```

Resumo modelo

```
summary(modelo_anova_2)
```

* REGRESSÃO

Pergunta 6:

Como o tempo de permanência é o fator crucial para pacientes internados, desejamos descobrir se o

tempo de permanência pode ser previsto a partir de idade, gênero e raça.

Resposta: Valor-p maior que 0.05 em todos os casos, logo, falhamos em rejeitar a H0.

O tempo de internação não pode ser previsto a partir das variáveis independentes usadas.

Usaremos um modelo de Regressão Linear.

Variável dependente: LOS (antes do ~)

Variáveis independentes: AGE, FEMALE e RACE (depois do ~)

H0: Não há relação linear entre variáveis dependente e independentes.

H1: Há relação linear entre variáveis dependente e independentes.

Criação modelo de regressão

```
modelo_lr <- lm(LOS ~ AGE + FEMALE + RACE, data = dados)
```

Resumo do modelo

```
summary(modelo_lr)
```

Pergunta 7:

Quais variáveis têm maior impacto nos custos de internação hospitalar?

Usaremos um modelo de Regressão Linear.

Variável dependente: TOTCHG (antes do ~)

Variáveis independentes: AGE, FEMALE, LOS, RACE e APRDRG (depois do ~)

```
names(dados)
```

H0: Não há relação linear entre variáveis dependente e independentes.

H1: Há relação linear entre variáveis dependente e independentes.

Criação modelo de regressão

```
modelo_lr_geral <- lm(TOTCHG ~ ., data = dados)
```

Resumo do modelo

```
summary(modelo_lr_geral)
```

Como podemos observar a partir dos valores dos coeficientes, as variáveis AGE(idade), LOS(tempo de permanência)

e APRDRG (grupo de diagnóstico refinado do paciente) têm três asteriscos (***) ao lado.

Então eles são os únicos com significância estatística;

Além disso, RACE não é significativo.

Vamos remover RACE e construir outra versão do modelo.

1º MUDANÇA

Criação do modelo de regressão sem a variavel RACE

modelo_lr_4var <- lm(TOTCHG ~ AGE + FEMALE + LOS + APRDRG, data = dados) summary(modelo_lr_4var)

Observe que a variável que representa o gênero tem a menor significância para o modelo.

Vamos removê-la e criar outra versão do modelo.

2º MUDANÇA

Criação do modelo de regressão sem a variavel FEMALE

modelo_lr_3var <- lm(TOTCHG ~ AGE + LOS + APRDRG, data = dados) summary(modelo_lr_3var)

As 3 variáveis tem alta significância, mas APRDRG tem valor t negativo.

Vamos removê-la e criar outra versão do modelo.

3º MUDANÇA

Criação do modelo de regressão sem a variavel APRDRG

modelo_lr_2var <- lm(TOTCHG ~ AGE + LOS, data = dados) summary(modelo_lr_2var)

A remoção de raça e gênero não altera o valor de R2.

A remoção do APRDRG no modelo aumenta o erro padrão.

Logo, o modelo modelo_lr_3var parece ser o melhor e o usaremos para nossa conclusão.

Melhor modelo

```
summary(modelo_lr_3var)
```

Conclusão:

Como é evidente nos vários modelos acima, os custos dos cuidados de saúde dependem

da idade, do tempo de permanência e do grupo de diagnóstico.

Essas são as 3 variáveis mais relevantes para explicar e prever o gasto com

internações hospitalares.