**Big Data e Cloud Computing - Assignment 2**

Sofia Malpique - up201704877

Yannik Bauer - up201805440

1. Descrição do problema

1. Packote para a resolução do problema

Decidimos usar a PySpark library para criar um dataframe, que editamos usando pipelines.

1. Analise estatistica e visualização

Primeiro escrevemos uma função que mostra os valores dos items com os itemids para um patiente e dia escolhido. Obtemos o seguinte gráfico:

Chart, scatter chart

Description automatically generated

A criação deste grafico demora mais ou menos 8 minutos.

Depois achamos interessante visualizar como é que os valores de um item ficam ao longo do tempo de um patiente. Obtemos por exemplo o seguinte gráfico, que demorou mais ou menos 10 minutos para criar.

Chart, line chart

Description automatically generated

Por último quisemos saber como é que é a distribuição do tempo que um patiente fica no hospital. O seguinte gráfico mostra quantos patientes ficaram quantos dias no hospital. Importante notar que o eixo de y não é linear. A criação deste gráfico demorou mais ou menos 10 minutos.

Chart, histogram

Description automatically generated

1. Predict length of stay