Seu pet center de estimação

Processo Seletivo Analista de Dados Jr.

Alexandre Edson

Relatório - Exercício 1

Introdução

Este relatório foi realizado para participação do processo seletivo da Petz onde será feita a análise dos dados de internações do SUS no período de dez/2017 até Jul/2019 incluindo as estimativas dos meses faltantes dentre esse período e a previsão de seis meses. Foi feito o Dashboard na plataforma da Microsoft Power BI e o tratamento dos dados utilizando a linguagem Python. Todo conteúdo é composto pelas argumentações representas em forma de gráficos.

Utilização da Linguagem Python

Conforme sugerido no case testes analista júnior, foi organizado/tratado as planilhas utilizando linguagem em Python. Foi desenvolvido três scripts.

O Primeiro script se refere ao tratamento dos dados do arquivo em Excel propriamente fornecida, onde ela é composta por 14 planilhas distribuídas aleatoriamente representando as periocidades dos dados por mês. O Script foi desenvolvido para fazer o tratamento de cada coluna identificando os tipos de dados, renomeação das colunas, remoção das colunas com dados faltantes (Sem interferir na futura análise), definir as casas decimais e fazer a limpeza da tabela deixando os dados padronizados e de forma mais compreendido. Identifiquei uma certa redundância quando encontrei o total de cada zona na mesma base de dados onde representa os estados. Para corrigir essa redundância, removi essas linhas com as zonas e inserir uma coluna representando a zona dos estados utilizando a API do IBGE para identificar a zona de cada Estado. A lógica foi feita de tal forma que é aplicada o mesmo tratamento e organização para cada planilha utilizando o loop for.

O Segundo script, foi desenvolvido para fazer a estimativa dos seis meses faltantes dentro do período fornecido. O processo foi utilizar a planilha gerada pelo script anterior, separar os dados por estado e fazer a média buscando os dois dados mais próximos da lacuna utilizando um dado referente ao mês anterior e um posterior a lacuna. Ao final é feita a união das planilhas em uma única incluindo a nova coluna “Período”.

O terceiro e último script é feita a previsão de meses a frente após o mês de julho de 2019 utilizando o modelo ARIMA. O Script utiliza a planilha gerada no script anterior, separa por estado e aplica o modelo para as colunas solicitas no case. Ao final é gerada a planilha somente com os dados dos meses previstos.

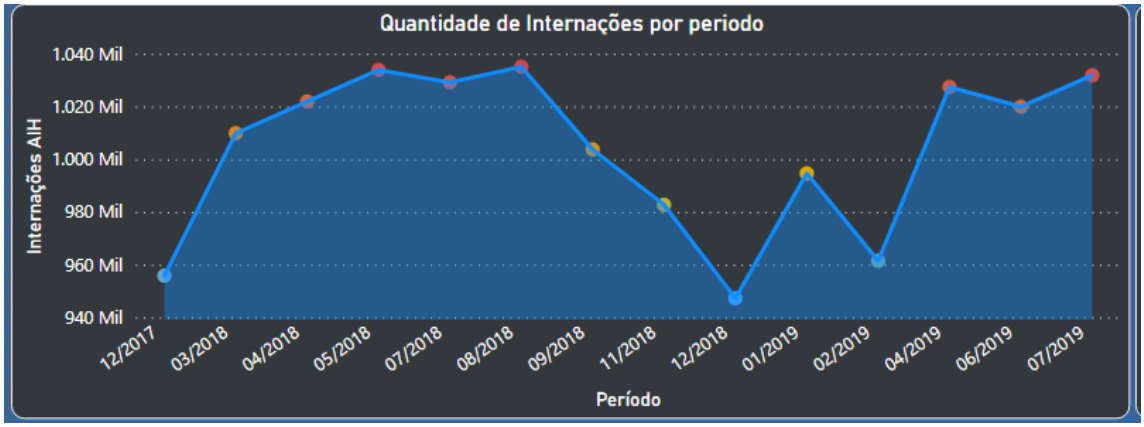
Os scripts estão disponibilizados em anexo a esse relatório no formato .html sendo possível visualizar em qualquer navegador e se encontra no OneDrive. Segue link abaixo:

https://1drv.ms/u/s!An1m2LCk81ZyjMw03utXqHr1VU8VHA?e=4IvMJ2

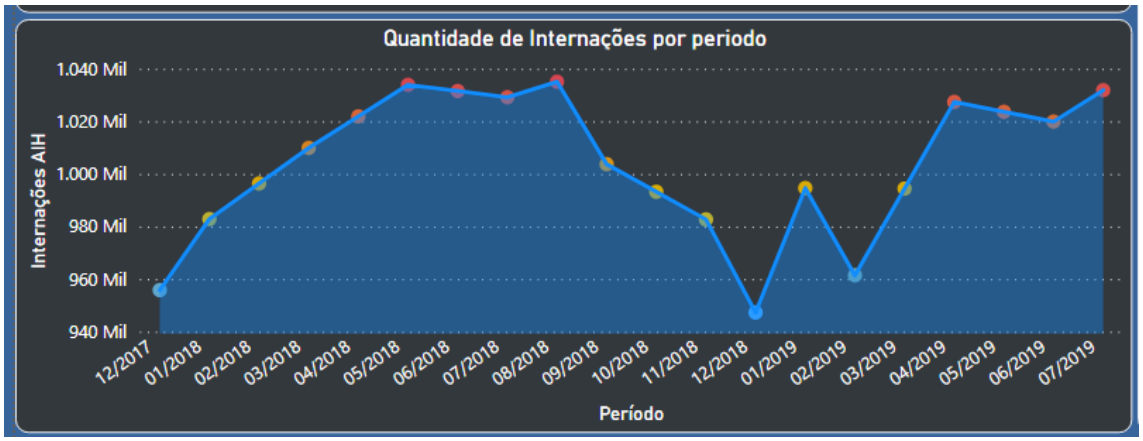
Análise dos dados

Para analisar os dados, criei um Dashboard na plataforma da Microsoft Power BI para fazer a representação das três planilhas geradas pelos scripts citado no item acima. No Dashboard é representado em três páginas, cada uma correspondente as planilhas citadas acima. Segue em abaixo o link do Dashboard:
https://app.powerbi.com/view?r=eyJrIjoiYjA0MDNkMTAtZjU3NC00YjUwLTkyMjUtYzdjYjQ1NjE5M2VlIiwidCI6IjU1ODg2NmRiLTY0YzMtNGZkMC05ZmViLWYzZGU0MDU0YmE1YSJ9

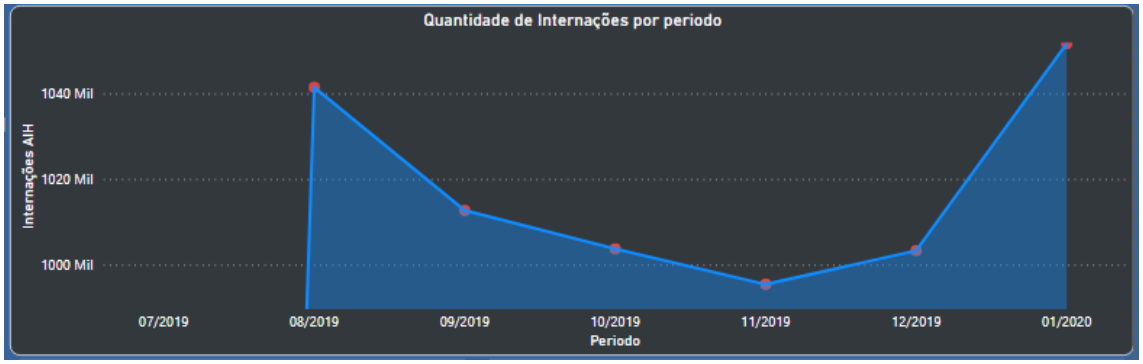
Quantidade de Internações por período



Dashboard Ex1 - Gráfico de área períodos incompletos



Dashboard Ex3 - Gráfico de área períodos completos



Dashboard Ex4 - Gráfico de área períodos previstos

Temos três gráficos de áreas representando a quantidade de internações AIH (Autorização de Internação Hospitalar). Foi escolhido esse tipo de gráfico para uma melhor visualização da diferencia de valores entre os períodos, que ao todo, durante os meses não demostram uma variam tão brutescas, uma vez que os valores variam entre 940 Mil e 1.040 Mil uma diferencia de 100 Mil internações ao decorrer dos meses.

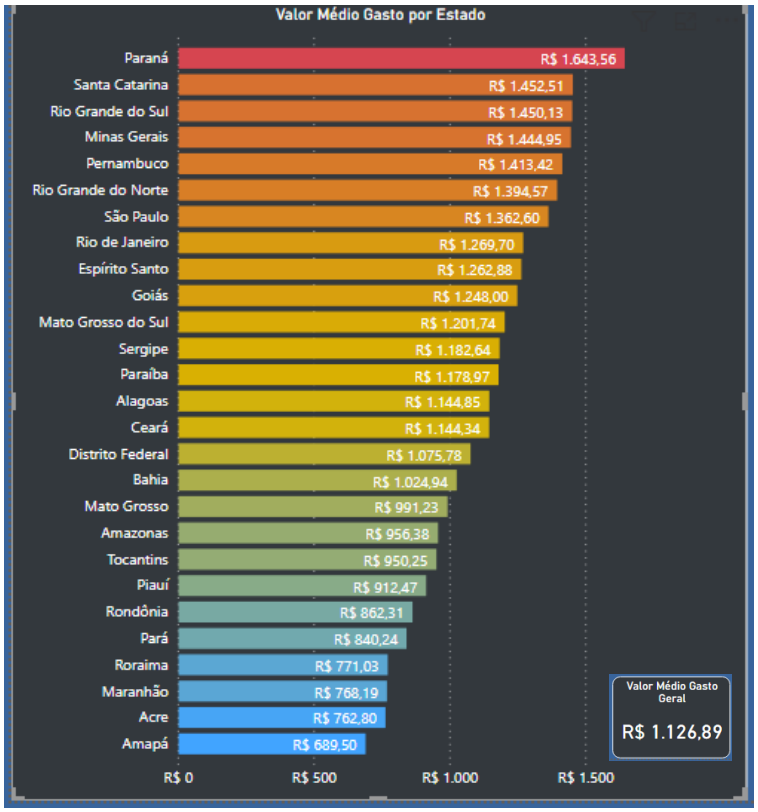
Analisando o **gráfico do Dashboard 03**, podemos ver que a estimativa não afetou nos valores dos meses existentes e seguindo a linha de tendencia do gráfico. Para estimar o valor dos meses faltantes podemos seguir com dois Métodos:

- Uma vez que não temos dados para fazer uma correlação com os dados independentes para estimar os dados dependentes, podemos usar os mais básicos que seriam os cálculos das Medias, Mediana e moda. Separar os dados por estado e aplicar os modelos.
- Outro método de estimar dados é usar o modelo KNN (K-Nearest Neighbours), esse método calcula a distância entre os valores mais próximos da lacuna onde se deseja fazer a imputação dos dados, uma vez feito isso, ele pegas esses valores e faz o cálculo da média.

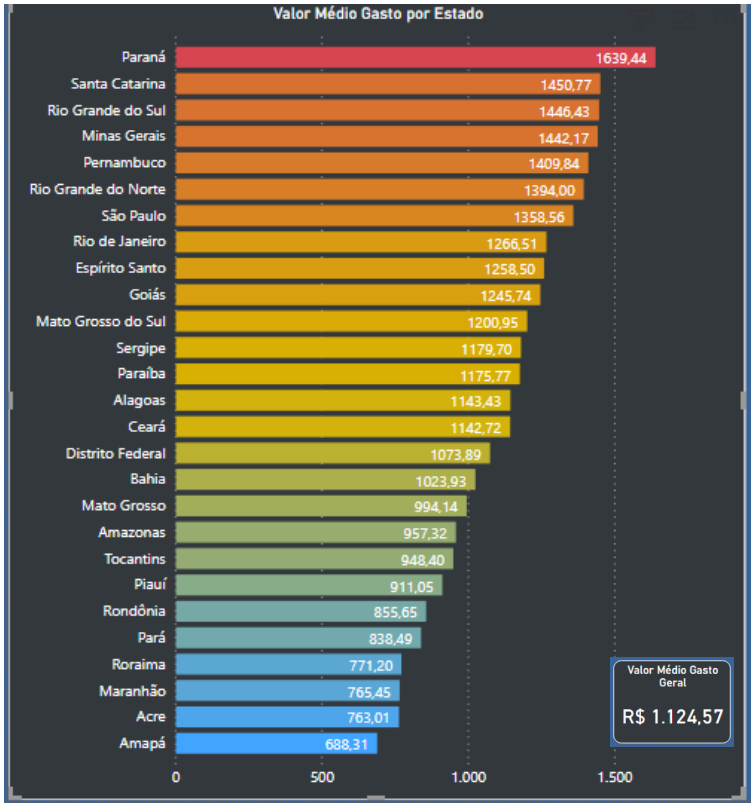
Para fazer a estimativa dos meses faltantes, Utilizei como base o método de KNN vizinhos mais próximos. Em python é feita a varredura dos dados faltantes buscando o valor posterior e anterior a cada lacuna. Após isso ele aplica a média para cada local vazio utilizando os próximos correspondentes.

O **gráfico do Dashboard 04** corresponde a previsão para os próximos seis meses. Utilizei o modelo ARIMA, pois estamos trabalhando com series temporais. Ele utiliza dados passados para prever o futuro usando os recursos de autocorrelação e média móveis. Como parâmetros de entrada é definido os parâmetros P, D e Q onde cada letra representa uma parte do nome ARIMA (P = AR, D = I, Q = MA) o P representa a Autorregressão, D Integração e o Q a média móvel. Podemos ver no gráfico que a previsão é que se mantenha em alta e tenha uma leve caída em forma de “U” e no mês de janeiro de 2020 a previsão diz que irar ultrapassar o maior pico que foi registrado em agosto de 2018, ultrapassando a quantidade de 1.040 Mil de Internados.

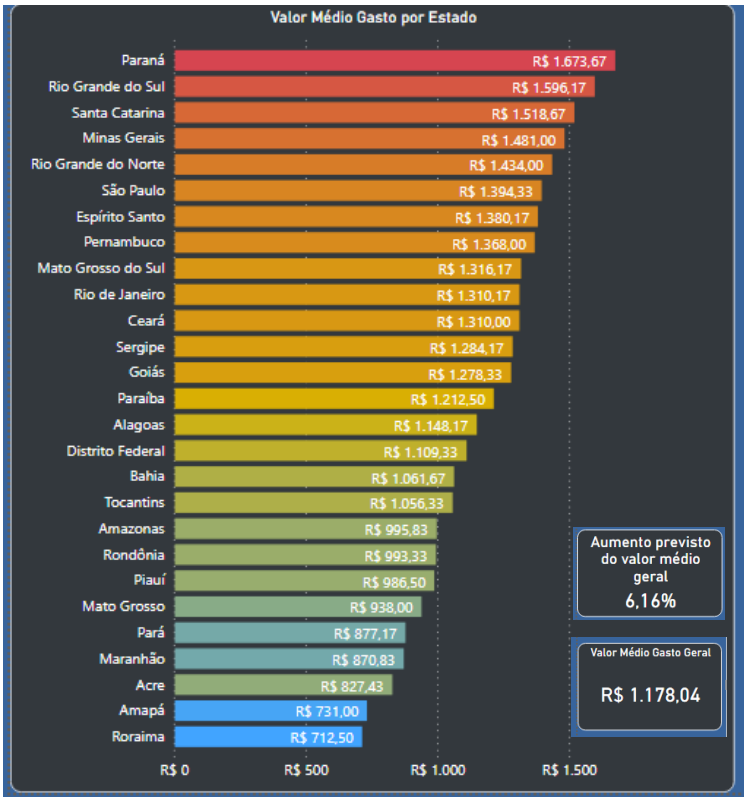
Valor Médio gasto por Estado



Dashboard Ex1 - Gráfico de barra períodos faltantes



Dashboard Ex3 - Gráfico de barra períodos completos



Dashboard Ex4 - Gráfico de barra períodos previstos

Na sequência acima temos três gráficos de barras para representar os valores totais gastos por estado equivalente as internações, classificados os estados com maior gasto com tendencia para o vermelho ou o estado com o menor gasto com tendência para a cor mais azulada. Foi escolhido gráfico em barra invertida devido a quantidade de estados serem bem considerados e para visualizar o nome do estado com mais clareza. Analisando os gráficos dos **Dashboard Ex1 e Ex3** não houve alterações entre eles. Podemos ver que existe uma diferencia de valores gasto por internações que devemos levar em conta. Por exemplo: pegando os estados das duas extremidades, podemos ver que o estado do Paraná tem um gasto mais que o dobro em relação ao estado de Roraima, então podemos levantar o seguinte questionamento: Qual foi o fator que levou ao estado do Paraná ter um gasto médio muito maior em relação ao outros estados? Será que seria mais viável buscar de uma distribuição buscando a equidade, o justo para cada estado, uma vez que o estado do Paraná não possui o maior número de Internações. Investigado a fundo o setor de distribuição de verbas para os estados, podemos diminuir a média gasta total que é por volta de 1.130,00 R$ e preservar o justo para cada estado. Para a previsão de durante os 6 meses a tendencia é o aumento de 6% em relação ao valor médio do Dashboard Ex3 permanecendo o padrão de diferencia do valor gasto por internação.

Total de internação por estado

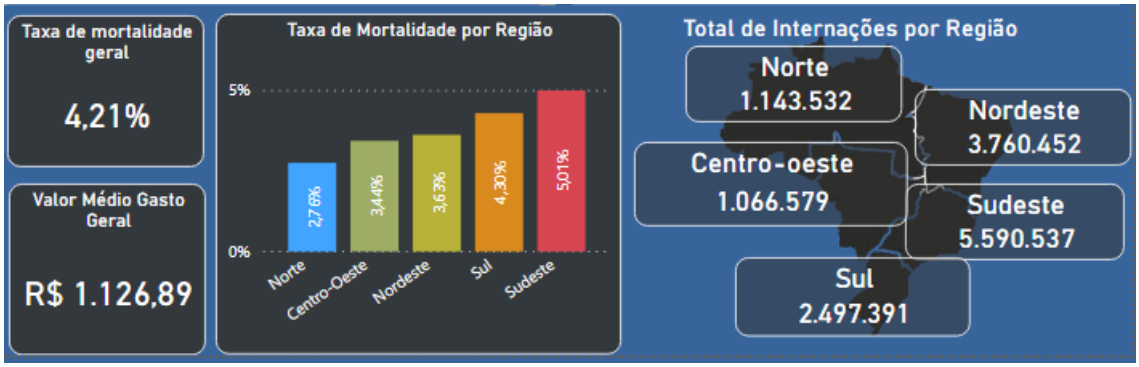
Dashboard Ex1- Gráfico de Mapa períodos faltantes

Dashboard Ex3 - Gráfico de Mapa períodos completos

Dashboard Ex4 - Gráfico de Mapa períodos previstos

A fim de buscar uma interação mais agradável, selecionei o mapa coroplético para representar o total de internações por estado, sendo que a coloração do mapa corresponde ao número de internações onde a cor mais voltada para o vermelho representa o maior número de internação, a cor voltada mais para o azul claro corresponde com o menor número de internações, consequentemente a cor próxima ao amarelo refere-se a mediana. É notório que as cores mais quentes estão localizadas mais próximas da região para o sudeste para o sul, locais com as temperaturas mais baixas, principalmente a região sul, já na região sudeste, está localizado em um ponto onde as clima tem bastante variações. Podemos concluir que esses lugares são mais propicio a procura de hospitais por doenças dos tipos respiratórias e gripais.

Análise por região



Dashboard Ex1 – Análise por região períodos previstos

Representação da taxa de mortalidade dividida por região onde temos a maior taxa em 5,01% representada pelo Sudeste, e a menor no Norte com 2,76%. Levando em conta a quantidade de internações no Brasil ser muito elevada dentro de um período de 14 meses, os óbitos de internados no brasil não passam de 5%.

Planejamento estratégico

Para diminuir o número de internações no Brasil, sugeria o controle melhor de autorização de internação com base em classificação de prioridades de atendimento, prioridades do tipo Acidentes, cirurgias etc. Para o estado de são Paulo, investiria em equipamento hospitalares para ocasionando um atendimento mais apropriando, e com recursos de tal forma que iniba a possibilidade de internar o paciente.