Relatório de Acesso dos Laboratórios de Ciência da Computação

Última atualização 14 de outubro de 2017

Amanda Vivian Alves de Luna e Costa
Lívia Cavalcanti Bandeira Julião

~~(objetivos iniciais procurar grupos de alunos que têm comportamento comum, verificar picos de acesso de dias da semana e de horas, cluster's de máquinas utilizadas, verificar que recursos que podem ser alocados ,de acordo com o usuário).~~

Os dados fornecidos foram os l*ogs* referentes ao mês de agosto de 2017, que possuem informações tais como: mês, dia, hora, máquina, status da máquina e identificação do usuário.

Para realizar as análises , foi utilizada a linguagem R e a IDE RStudio.

Inicialmente, organizamos os dados em *data frames*, em que cada coluna representava um tipo de informação, após isto, filtramos estes dados para apenas as sessões abertas, e demos início às análises.

Os primeiros aspectos analisados foram: a média de acessos do mês ,a partir do dia da semana; os acessos por turno; qual laboratório é mais utilizado e quais os intervalos de tempo em que há mais acessos (desconsiderando os minutos e segundos); também, quais usuários mais *logaram* e quais as máquinas mais utilizadas pelos usuários.

Estas análises possibilitaram a construção de gráficos de barras, que foram plotados no RStudio e serão apresentados posteriormente.

1. Qual a quantidade de acessos diários dos computadores ?

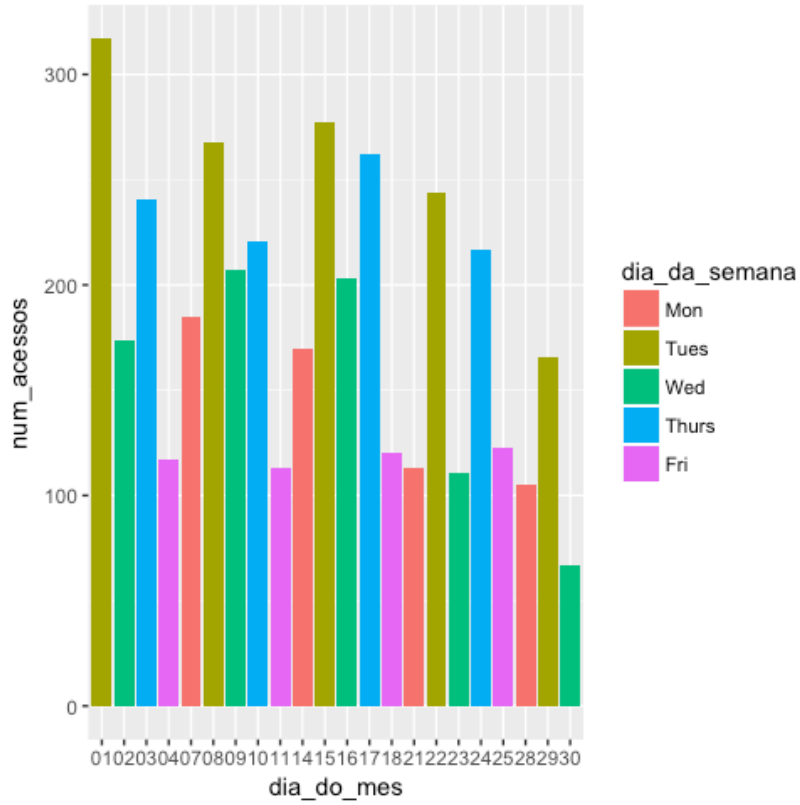


Gráfico 01 - mostra a quantidade de acessos por dia.

Verifica-se, a partir do gráfico acima, que, no primeiro dia do mês, foi quando ocorreu o maior número de acessos, verifica-se também que usualmente a quantidade destes acessos é superior a 100.

2. Quais os dias da semana em que ocorrem mais acessos?

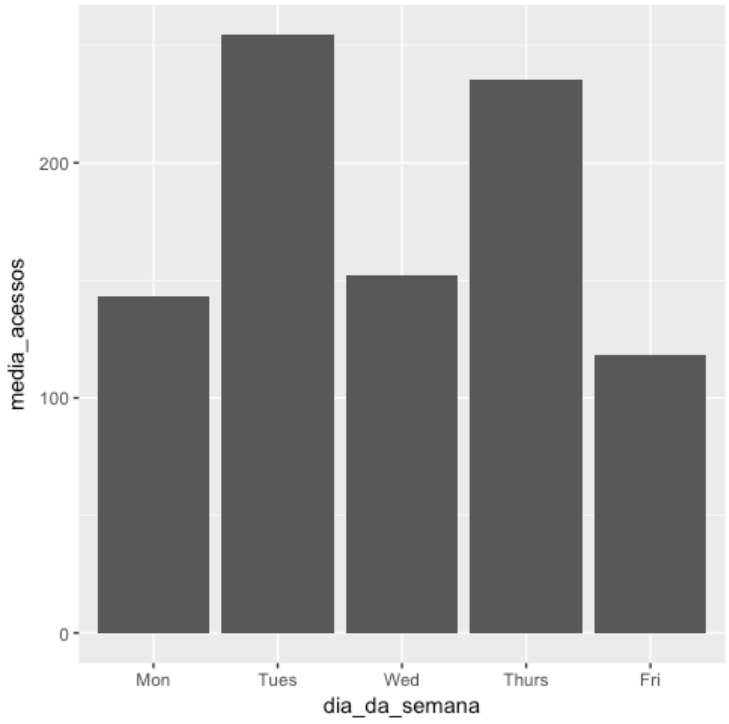


Gráfico 02 - Mostra a quantidade de acessos por dia da semana.

Vê-se ,pelo gráfico, que a terça-feira é o dia da semana que mais ocorrem acessos às máquinas, seguida da quinta-feira, e da sexta-feira quando há menos utilização dos laboratórios, porém o uso dos alunos para a realização dos minitestes de programação I ,que ocorrem neste dia não são contabilizadas, por utilizarem uma imagem diferente da padrão, fazendo com que haja uma perda de precisão ao analisar. A maior quantidade da terça e quinta podem ser explicadas devido a maior quantidade de aulas de programação serem nestes dias.

3. Quais os horários quando há mais acessos?

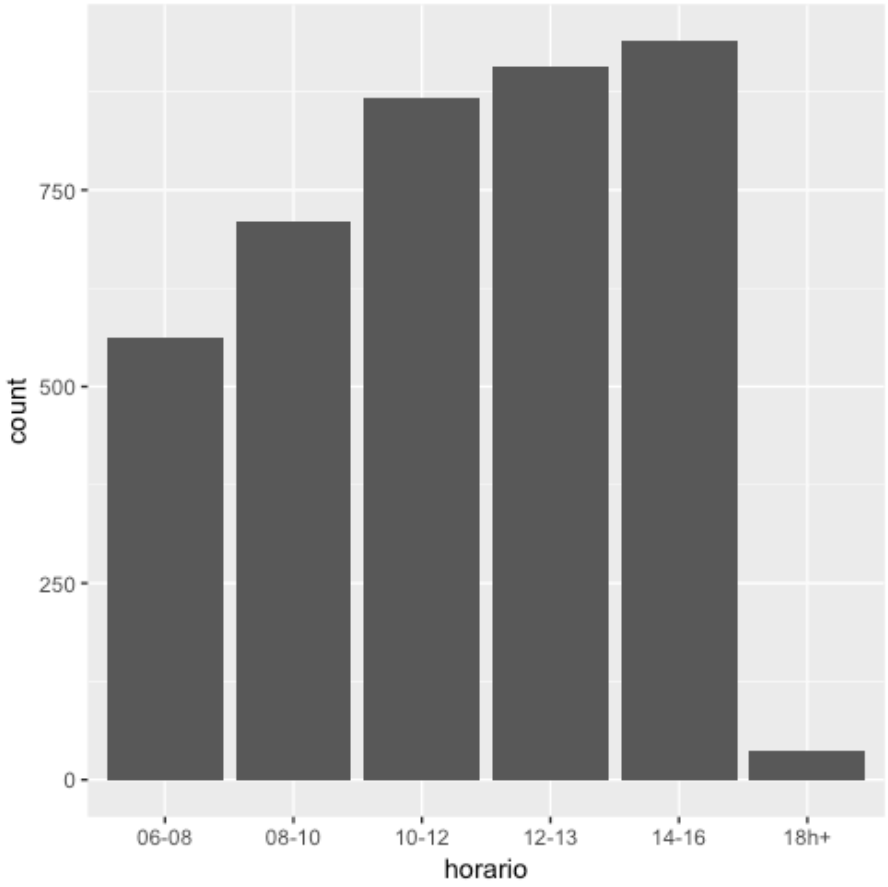


Gráfico 03 - mostra a quantidade de acessos às máquinas em blocos de duas horas.

Observa-se ,pelo gráfico, que os horários em que ocorrem mais acessos estão no intervalo das 14h às 16h, diferente do que era esperado(12h às 13h), pois geralmente este horário é dedicado a aulas e os computadores são ocupados apenas pelos estudantes daquela determinada disciplina e o horário das 12h às 13h é um horário no qual estudantes de todos os períodos possuem acesso ás máquinas, o que em tese faria com que o número de acessos neste horário fosse maior

4.Qual laboratório é mais utilizado?

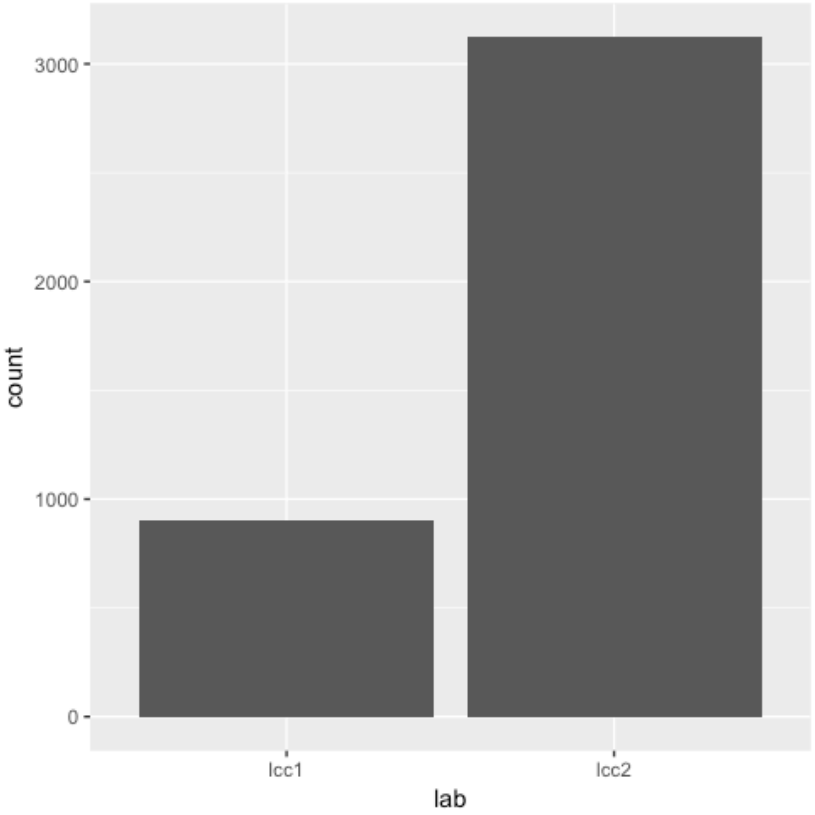


Gráfico 04 - mostra qual laboratório foi mais utilizado.

O gráfico mostra que o LCC2 foi o laboratório mais utilizado pelos alunos.
(Por que??)

5.Qual a quantidade de acessos por turno?

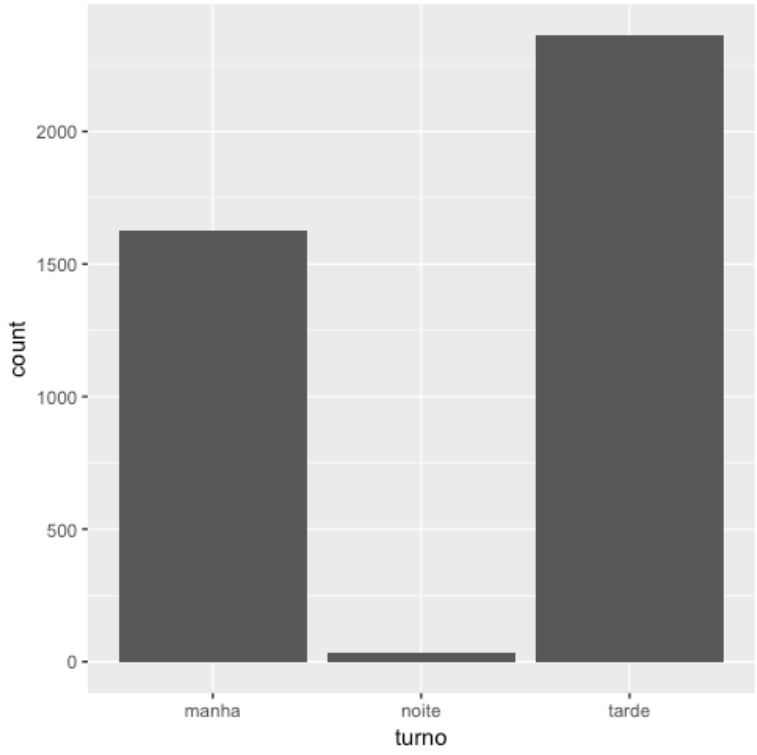


Gráfico 05 - mostra a média de acessos por turno.

Verifica-se que o turno da tarde, ao contrário do que se esperava(por que?), o turno da tarde é o que ocorrem mais acessos aos laboratórios.

6.Clusterização

Com os dados tratados, foi-se em busca de padrões de acesso tanto de usuário quanto de máquina.

Inicialmente, pensou-se em clusterizar a população baseado em dois atributos: máquina e usuário, individualmente. Para tal, verificou-se a quantidade ideal de grupos através da silhouette analysis, que é utilizada para estudar a similaridade de cada ponto a um cluster e, então, avaliar, por exemplo, a quantidade ideal de clusters.

Foram feitas análises de com o método "silhouette" para grupos de 2 a 10 e os resultados foram guardados numa matriz de dissimilaridade. Dissimilaridade pode ser definida como a distância entre dois pontos, que num plano Cartesiano seria uma distância Euclidiana.

A partir dessa matriz foi obtido um gráfico de linha que se mostra, desde o começo, descrecente. Portanto o valor ideal de clusters é a primeira coordenada do eixo x, ou seja, dois. Caso o gráfico fosse crescente, o valor ideal seria imediatamente antes de ele começar a decrescer, ponto no qual o agrupamento de torna gradativamente mais forte chegando ao grupo de um ponto.

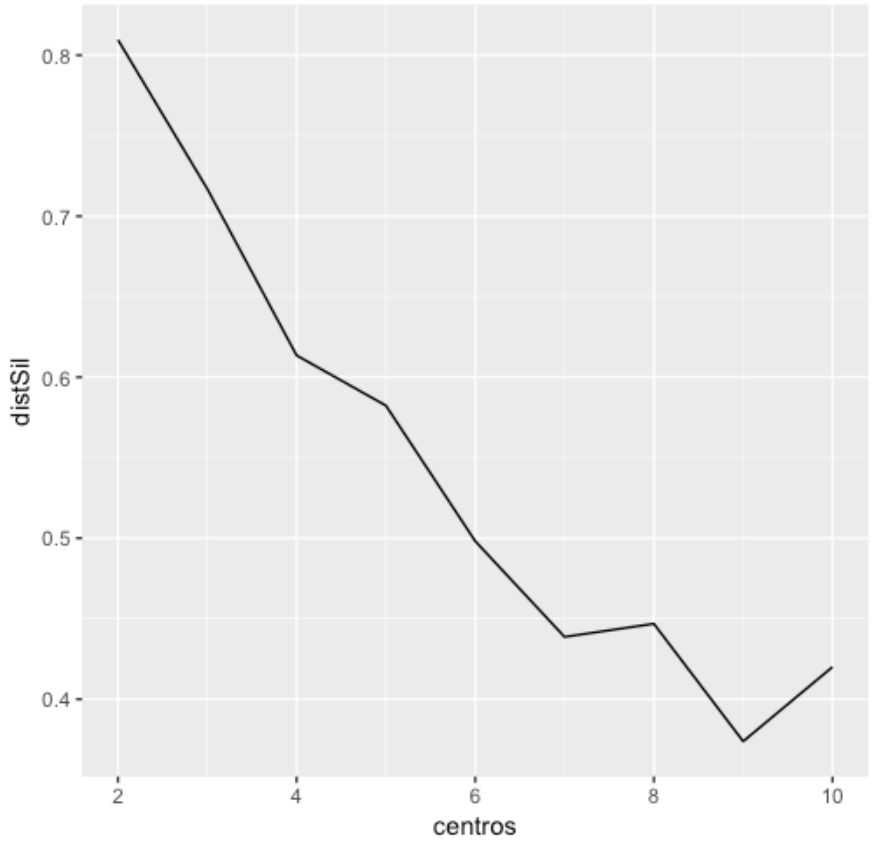
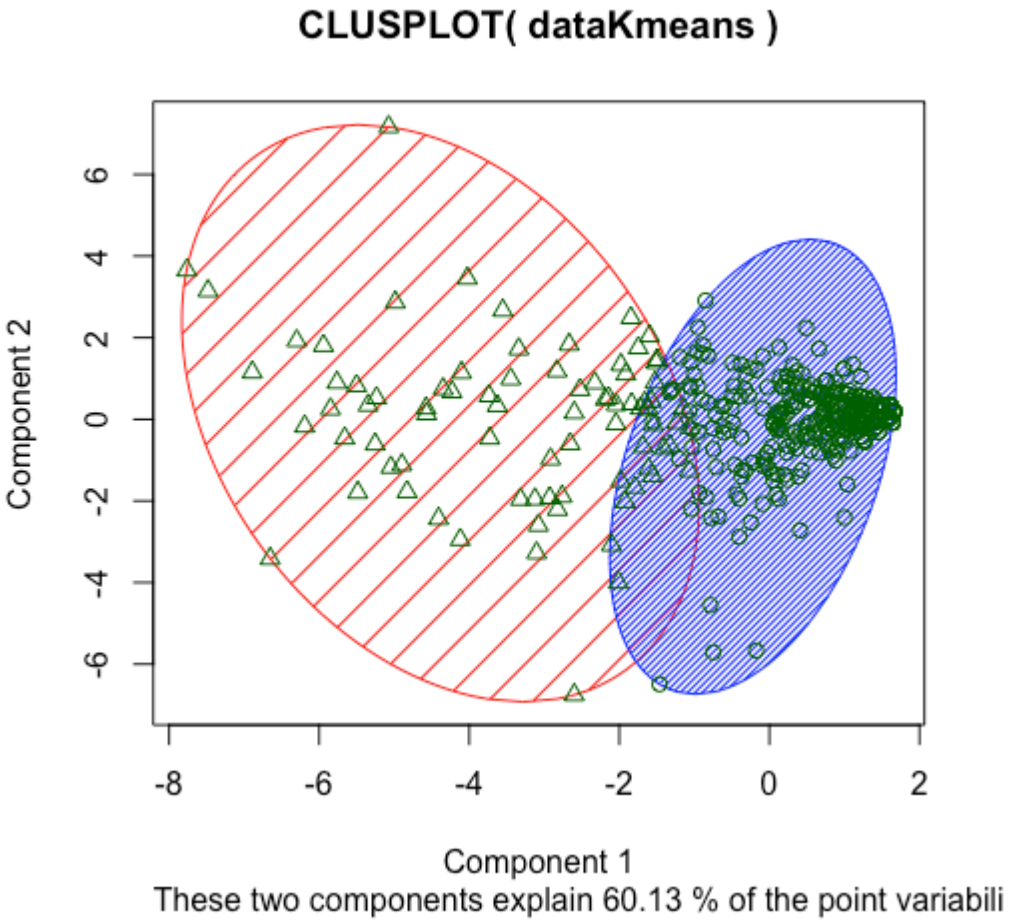


Gráfico 06 - Grafico do silhouette

Com base nisso, o k-means foi utilizado por ser um algoritmo simples e muito conhecido para clusterização.



Além do gráfico, tais números foram estabelecidos:

Centro	Número de acessos	6h-8h	8h-10h	10h-12h	12h-14h	14h-16h	16h-18h	18h+	Número de pontos
1	24.666667	0.19753086	4.012346	4.716049	9.938272	2.3950617	3.2222222	0.18518519	81
2	5.323684	0.08421053	1.013158	1.273684	1.713158	0.5157895	0.6684211	0.05526316	380