**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

* DADOS E FERRAMENTAS

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.

* ANÁLISE

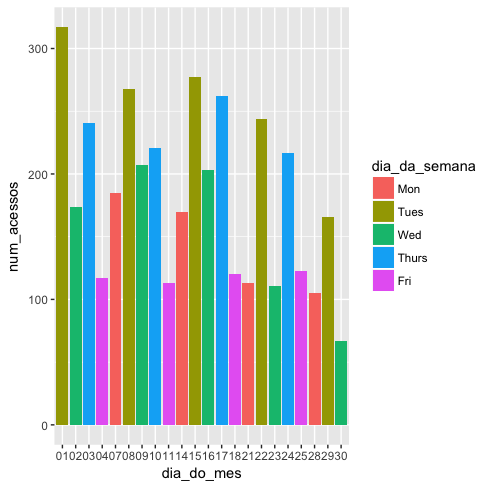
Inicialmente, organizamos os dados em *data frames*, em que cada coluna representava um tipo de informação, após isto, filtramos estes dados para apenas as sessões abertas, e demos início às análises.

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.

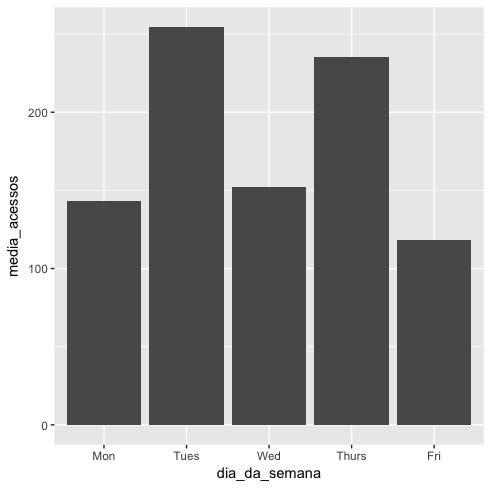
* PERGUNTAS

**1. Qual a quantidade de acessos diários dos computadores ?**

  
Gráfico 01: mostra a quantidade de acessos por dia.

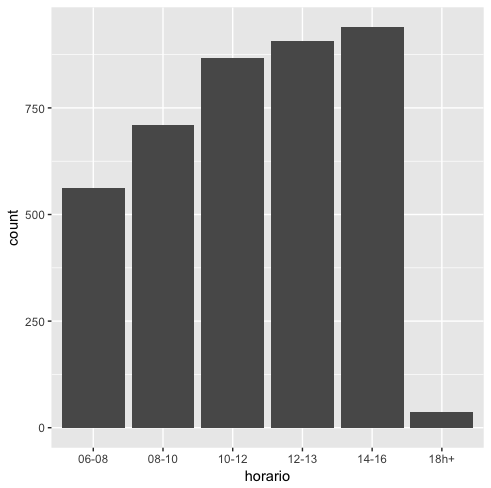
Verifica-se, a partir do gráfico ao lado, que, no primeiro dia do mês, foi quando ocorreu o maior número de acessos, verifica-se também que usualmente a quantidade destes acessos é superior a 100.

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

****

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.

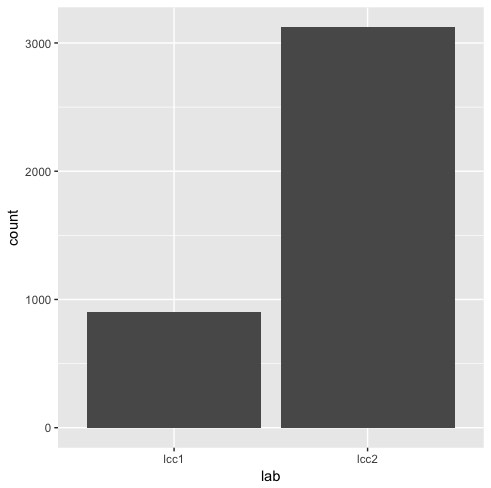
**3. Quais os horários quando há mais acessos?**

****

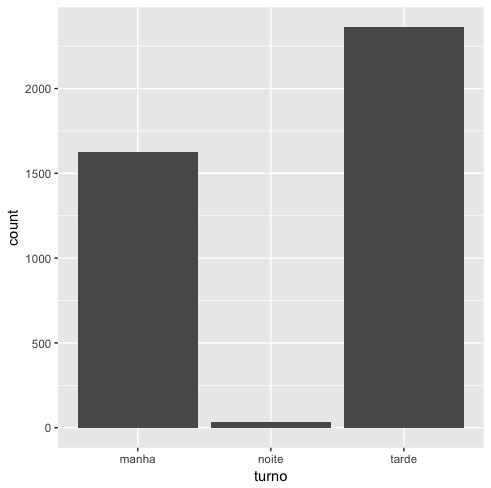
Observa-se, pelo gráfico, que os horários em que ocorrem mais acessos estão no intervalo das 14h às 16h, diferente do que era esperado(12h às 13h), pois geralmente este horário é dedicado a aulas e os computadores são ocupados apenas pelos estudantes daquela determinada disciplina e o horário das 12h 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?**

O gráfico mostra que o LCC2 foi o laboratório mais utilizado pelos alunos. Possivelmente pela maior disponibilidade que corresponde ao horário de aulas, 08-12 hrs e 14-20 hrs, e o horário de almoço, 12-14 hrs. Enquanto que o LCC1 fica, geralmente, com o uso restristo às aulas e às atividades dos guardians.

  
Gráfico 04 - mostra qual laboratório foi mais utilizado.

**5.Qual a quantidade de acessos por turno?**

****

Verifica-se que no turno da tarde, ao contrário do que se esperava, ocorrem mais acessos aos laboratórios. As possiblidades que podem explicar os resultados são (a) o horário de acesso aos LCC’s considerado manhã são de 05 às 11 hrs, sendo que o horário inicial dos acessos é, geralmente, de 08 hrs – hora de início das atividades acadêmicas na instituição, na maioria das vezes - e o da tarde de 12 às 17 hrs, logo o tempo de acesso considerado tarde é maior do que de manhã o que leva a mera contagem de número de acessos a uma análise enviesada, sendo preferível o uso de outras métricas para melhor análise, talvez; (b) como os alunos podem ficar mais ociosos no horário da tarde, por englobar o período de almoço e de espera para aulas da tarde – que geralmente fazem com que os estudantes fiquem mais na universidade e, com frequência, no LCC2 – a quantidade de acessos nesse turno é incrementada.

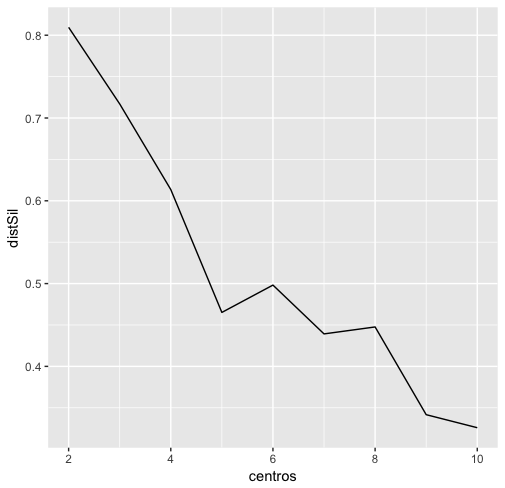
**6.Clusterização**

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

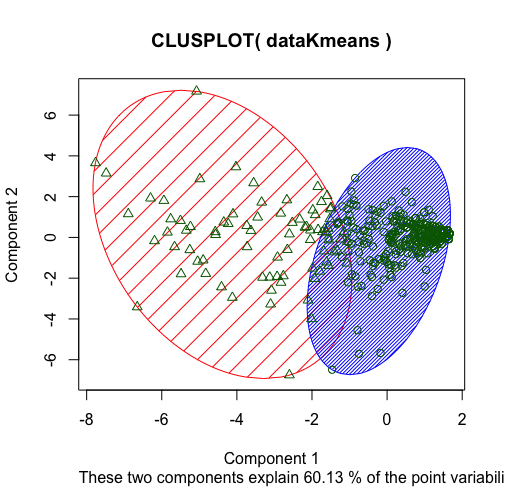
Inicialmente, pensou-se em clusterizar a população baseado em dois atributos: máquina e usuário, individualmente. Para tal, verificou-se a quantidade ideal de grupos através da silhouette analysis, que é utilizada para estudar a similaridade de cada ponto a um cluster e, então, avaliar, por exemplo,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

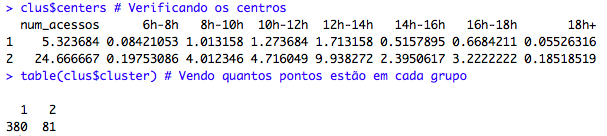
A partir dessa matriz foi obtido um gráfico de linha que se mostra, desde o começo, decrescente. Portanto o valor ideal de clusters é a primeira coordenada do eixo x, ou seja, dois. Caso o gráfico fosse crescente, o valor ideal seria imediatamente antes de ele começar a decrescer, ponto no qual o agrupamento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



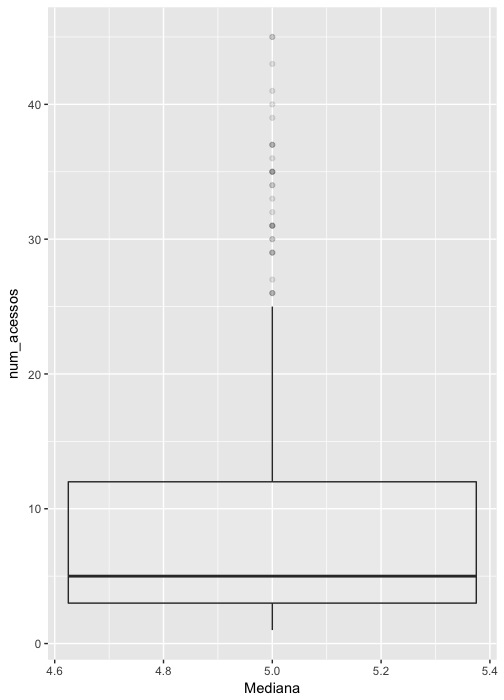
Com base nisso, o k-means foi utilizado por ser um algoritmo simples e muito conhecido para clusterização.



Além do gráfico, tais números foram estabelecidos:



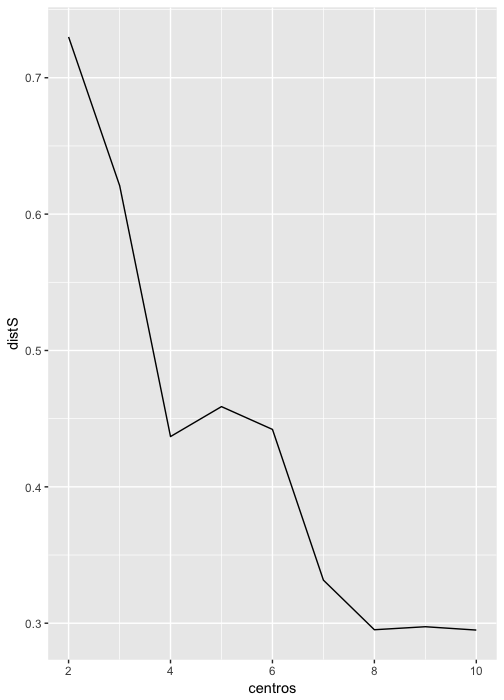
Dado os resultados do k-means, outro estudo dos dados foi feito a partir de um boxplot a fim de analisar-se a distribuição do número de acessos.

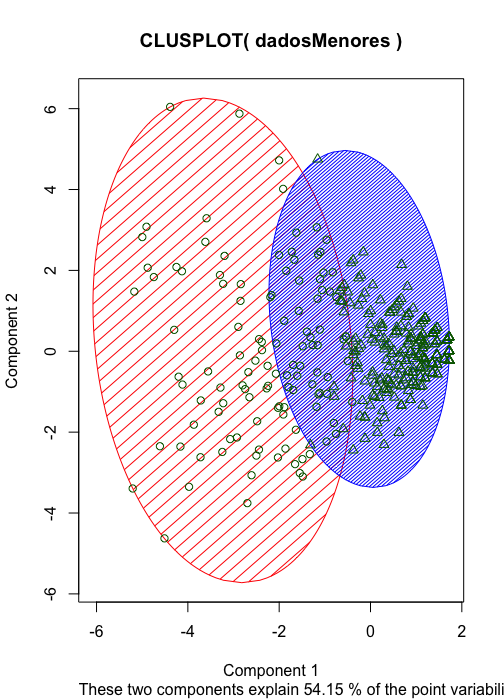
  
Gráfico 08: boxplot do número de acessos

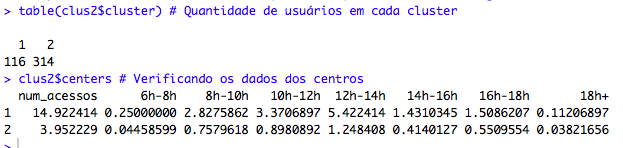
Nele vê-se que o menor valor é 1, já que um aluno só é contabilizado se logar pelo menos uma vez no período de estudo. O número máximo de acessos é 45 e está representado pelo último círculo de baixo para cima. O retângulo é formado pelo primeiro quartil, linha inferior, que, nesse caso, é cinco e deixando 25% dos dados abaixo. O retângulo grande é fechado pela linha do terceiro quartil que marca 75% do número de acessos abaixo de si e vale 12. A linha horizontal dentro da figura retangular considerada é a mediana que divide os dados ao meio e é igual a 5. Sua posição define que os dados são positivamente assimétricos, ou seja, metade dos acessos tem magnitude baixa em relação ao intervalo considerado. Outra possível interpretação  é que dado que o mês em estudo deve cinco semanas, a maior parte dos estudantes acessou sua conta até uma vez na semana. Também é possível observar que 50% dos acessos está entre 5 e 12. E que são considerados outliers os alunos com quantidade de log in entre 26 e 45. Logo, mais da metade utiliza os LCC’s mais que semanalmente, chegando a uma média de quase duas vezes ao dia.

No entanto, como é possível ver pela quantidade considerável de círculos no boxplot, os dados apresentam muitos outliers, que são indivíduos que residem fora do padrão de distribuição dos dados. No nosso caso, são considerados outliers os alunos que tiveram número de acessos maior que 25. Dado isso, foram filtradas as observações com até 25 acessos e calculado um novo k-means.

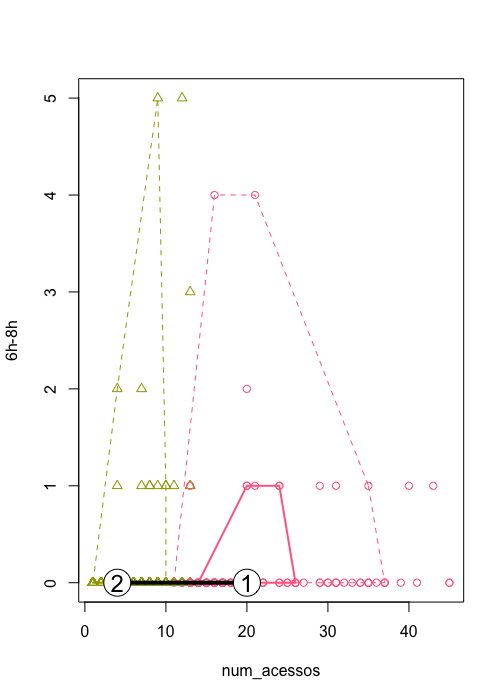
O número de clusters se manteve igual a dois e as diferenças entre os grupos diminuíram, como se pode ver na tabela.

  
Gráfico 09: Silhouette dos dados com até 25 acessos

  
Gráfico 10: Gráfico do K-means dos dados com até 25 acessos



Para que a nossa análise de clusterização a partir do k-means não ficasse prejudicada por causa dos outliers, escolheu-se outro método, o K-median, que é menos prejudicado por valores extremos devido ao uso da mediana para os cálculos. Além disso, como esse método escolhe valores que, necessariamente, estão nos dados é possível identificar os elementos mais representativos do grupo. Caso também seja utilizada a distância Euclidiana, como em sua definição formal, ao contrário da distância Euclidiana quadrática usada no k-means, o modelo de k-mediana se torna mais robusto para elementos isolados presentes nos dados.

  
Gráfico 11: k-medians dos dados