Kaggle: Bike Trips

Romario

2022-12-15

Introdução

Serem acusados de fazerem mal ao meio-ambiente e terem um aumento, quase logarítmico, dos custos com impostos e manutenção, tornaram os carros cada vez mais impopulares.

Em contrapartida, vem aumentando o uso de aplicativos de mobilidade, individual e coletiva. Tanto de automóveis, como de patinetes e bicicletas.

Aqui no Rio, cidade onde moro, houve um boom dos patinetes, no mundo distante do final década e 10. Hoje,devido a motivos sanitários e políticos, talvez, eles perderam força

Outro meio de locomoção alternativo, iniciado antes dos patinetes e que ainda sobrevivem, são as bicicletas alugadas. Elas são fornecidas por um grande popular no país, não citarei nomes, pois não estou sendo patrocinado.

Não sei os custos para alugá-las, pois não sei andar de bicicleta, mas é possível ver pessoas felizes as pedalando nas orlas da zona sul. Seja alguém que gosta de se exercitar ou um entregador de deliveries.

A ideia é boa e parece lucrativa, ou pelo menos dá uma aura de sustentabilidade à empresa. Entretanto, não é original, o que não é um problema. Ao que parece, a Citi Bike é um programa parecido de compartilhamento de bicicletas, presente na cidade de Nova Iorque desde de 2013.

Olha, minha visão de Nova Iork é um pouco estereotipada, influência de comédias românticas e gibis do homem-aranha, então não me parece ser uma cidade muito agradável e segura para andar de bicicleta.

Mas se os nova-iorquinos se aventuram, quem sou eu para julgar. E é sobre eles que vamos aprender mais um pouco.

Cadê as fontes?

Essa análise será feita a partir de um dataset disponível no Kaggle, uma acervo de vários arquivos para treinar:

- Análise de dados;
- Visualização de dados;
- Machine Learning;
- e muitas outras coisas.

Além disso, ele possui cursos com certificados de introdução a linguagem python e ao principal pacote do mesmo para a área de dados, o pandas.

Mas prefiro usar o R.

Este é o clique aqui para o `dataset`, bora a a análise.

Bibliotecas

Uma primeira visão dos dados

Vamos descobrir qual o tamanho deste dataset, seu número de linhas e colunas.

```
## [1] "Número de linhas: 1595334 || Número de variavéis: 11"
```

O primeiro número é de linhas e o segundo e colunas. E bem temos muitas linhas.

Devemos também visualizar se há células vazias, com linhas ou colunas com informações faltantes.

```
## [1] "Número de informações faltando 0"
```

Pelo santos guias do BI, esse dataset não possuí dados faltantes, obrigado Biel :D

Agora veremos as o nome de todas colunas

```
##  [1] "start_time"         "stop_time"          "start_station_id"
##  [4] "start_station_name" "end_station_id"     "end_station_name"
##  [7] "user_type"          "bike_id"            "gender"
## [10] "age"                "trip_duration"
```

Repare que há uma coluna determinada `bike_id`. No texto que acompanha o datase, eu havia entendido que era um id para8 *cada usuário* do aplicativo resolvi dar uma olhada.

bike_id	age	gender
25805	32	male
25805	44	male
25805	31	female
25805	31	female
25805	31	female
25805	27	male

Em bancos de dados ou estruturas para análise de dados, os id identificam algo único. Se ele fossem uma representação de um único usuário em gênero e idade, não mudariam em menos de o mês. O que me faz suspeitar, que este id representa uma bicicleta especídica.

No mundo real, numa realidade de dados, o ideal seria perguntar a quem faz o recolhimento desses dados, o que esse bike_id significa. Como não posso perguntar ao Biel, irei supor que dados sobre gênero e id, independem do bike_id. Essa suposição será importante para continuar a análise.

Terminada esta parte, vamos ver sobre o tipo desses dados.

```
## Rows: 1,595,334
## Columns: 11
## $ start_time         <dttm> 2018-05-31 23:59:59, 2018-05-31 23:59:59, 2018-05-~
## $ stop_time          <dttm> 2018-06-01 00:12:57, 2018-06-01 00:12:26, 2018-06-~
## $ start_station_id   <dbl> 312, 401, 483, 3107, 3341, 3562, 479, 128, 537, 322~
## $ start_station_name <chr> "Allen St & Stanton St", "Allen St & Rivington St",~
## $ end_station_id     <dbl> 460, 360, 368, 3076, 3400, 3562, 3635, 308, 546, 33~
## $ end_station_name   <chr> "S 4 St & Wythe Ave", "William St & Pine St", "Carm~
```

```
## $ user_type          <chr> "Subscriber", "Subscriber", "Subscriber", "Subscrib~
## $ bike_id            <dbl> 25805, 17258, 19692, 28285, 21000, 32205, 31327, 20~
## $ gender             <chr> "male", "male", "male", "male", "female", "male", "~
## $ age                <dbl> 32, 24, 39, 28, 51, 25, 34, 32, 37, 48, 31, 27, 26,~
## $ trip_duration      <dbl> 12.97, 12.45, 8.28, 7.75, 8.05, 16.43, 3.15, 10.52,~
```

Nada a comentar.

Vamos resumir esses dados?

```
##                    vars       n     mean      sd      min      max    range
## start_time            1 1595334      NaN      NA      Inf     -Inf     -Inf
## stop_time             2 1595334      NaN      NA      Inf     -Inf     -Inf
## start_station_id      3 1595334  1548.19 1427.93    72.00   3686.0   3614.0
## start_station_name    4 1595334      NaN      NA      Inf     -Inf     -Inf
## end_station_id        5 1595334  1537.10 1426.69    72.00   3686.0   3614.0
## end_station_name      6 1595334      NaN      NA      Inf     -Inf     -Inf
## user_type             7 1595334      NaN      NA      Inf     -Inf     -Inf
## bike_id               8 1595334 26201.61 5784.64 14529.00  33690.0  19161.0
## gender                9 1595334      NaN      NA      Inf     -Inf     -Inf
## age                  10 1595334    37.86   11.03    16.00     65.0     49.0
## trip_duration        11 1595334    16.43  284.86     1.02 111781.7 111780.7
##                      se
## start_time           NA
## stop_time            NA
## start_station_id   1.13
## start_station_name   NA
## end_station_id     1.13
## end_station_name     NA
## user_type            NA
## bike_id            4.58
## gender               NA
## age                0.01
## trip_duration      0.23
```

HUm... Enganei-me. As colunas end_station_id, start_station_id e bike_id apesar de serem números, acredito que serem caracteres seja melhor, para evitar algum futuro erro.

```
##                    vars       n  mean     sd   min      max    range   se
## start_time            1 1595334   NaN     NA   Inf     -Inf     -Inf   NA
## stop_time             2 1595334   NaN     NA   Inf     -Inf     -Inf   NA
## start_station_id      3 1595334   NaN     NA   Inf     -Inf     -Inf   NA
## start_station_name    4 1595334   NaN     NA   Inf     -Inf     -Inf   NA
## end_station_id        5 1595334   NaN     NA   Inf     -Inf     -Inf   NA
## end_station_name      6 1595334   NaN     NA   Inf     -Inf     -Inf   NA
## user_type             7 1595334   NaN     NA   Inf     -Inf     -Inf   NA
## bike_id               8 1595334   NaN     NA   Inf     -Inf     -Inf   NA
## gender                9 1595334   NaN     NA   Inf     -Inf     -Inf   NA
## age                  10 1595334 37.86  11.03 16.00     65.0     49.0 0.01
## trip_duration        11 1595334 16.43 284.86  1.02 111781.7 111780.7 0.23
```

Acho que assim está melhor

Vamos conhecer o público?

Acho que a primeira pergunta é.... Quem usa essas bicicletas? Isso é importante para a empresa pensar em que como fazer campanhas, tanto para agradar seu publico mais fiel ou para atrair pessoas novas.

Ou, em casos de verba para melhor o serviço, que tipo de público-alvo ou persona ela deve focar em agradar, para manter a fidelidade.

Então a primeira pergunta é:

Qual o gênero mais usa o serviço?

Gênero	Quantidade	Porcentagem
Feminino	415577	0.26
Masculino	1179757	0.74

Claramente, pessoas do gênero masculino são as que mais usam o produto.

Qual idade?

Precisamos categorizar a idade, para tornar mais fácil a visualização e entendiento

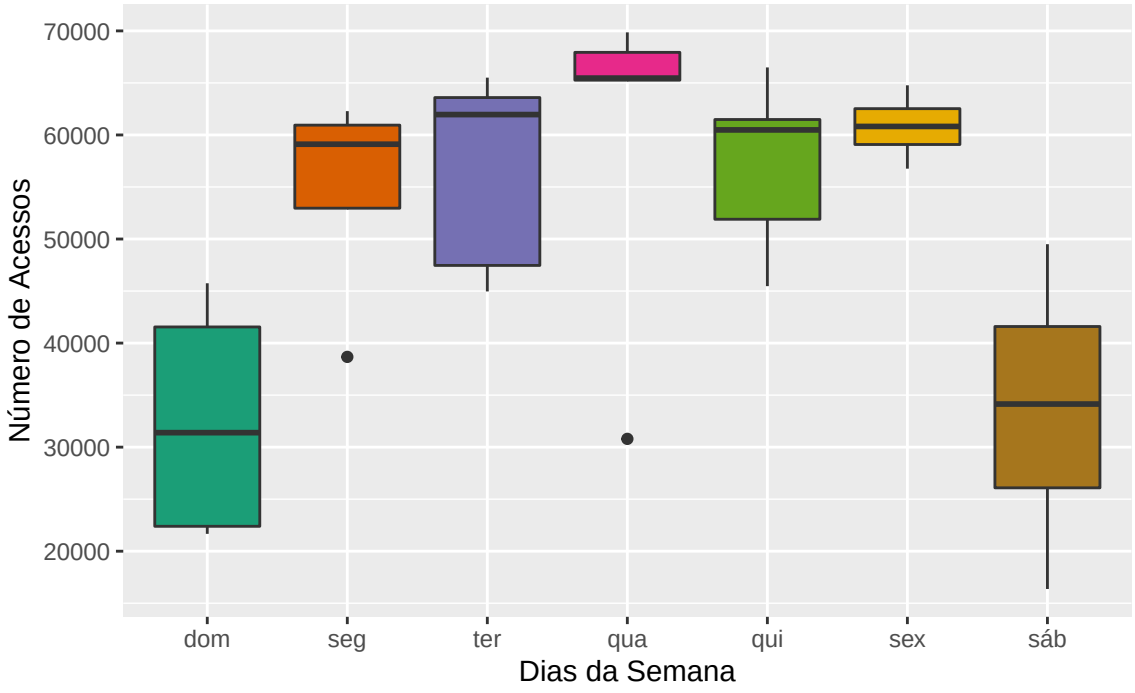
Faixa Etária	Quantidade	Porcentagem
0-19	21282	0.01
20-29	481418	0.30
30-39	517432	0.32
40-49	313132	0.20
50-59	209923	0.13
60-69	52147	0.03
70-79	0	0.00
acima de 80	0	0.00

O maior público das bicicletas está está na faxa dos 20 à 39 anos, mas da metade dos usuários.

Visualizações

Esses dados são do mês de Maio de 2018, não sei se esse mês é possuí alguma data especial em Nova Iorque, mas será que ele pode dar informação não só sobre o marketing, mas também sobre a logística?

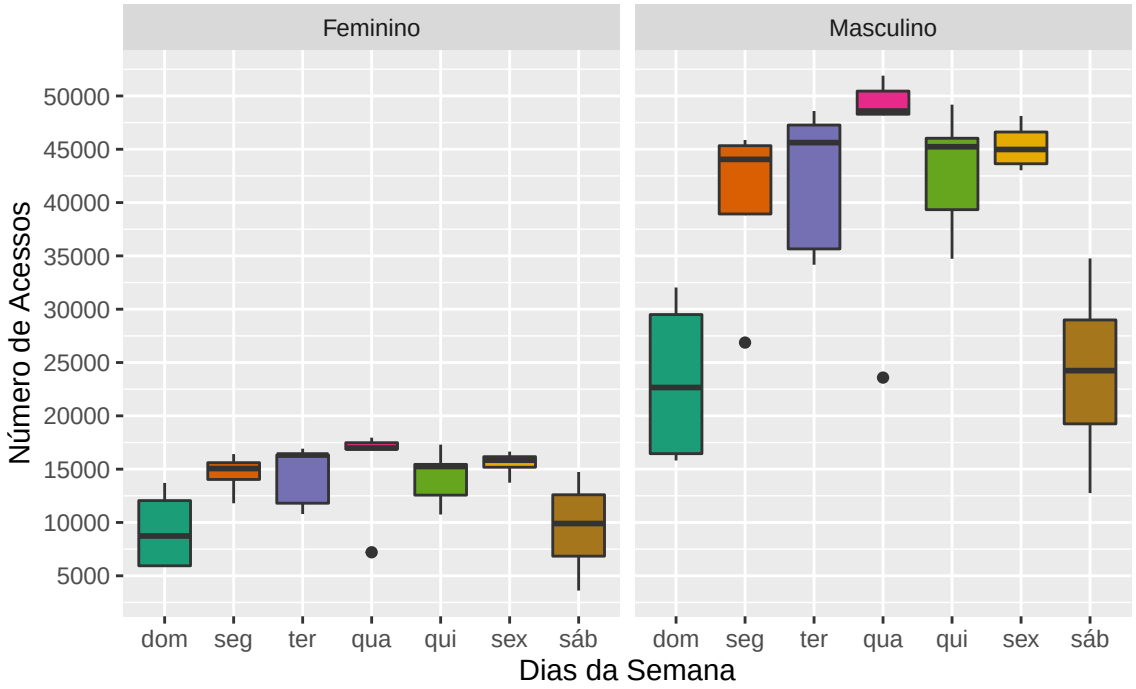
Quais dias da semana tem mais usuário?



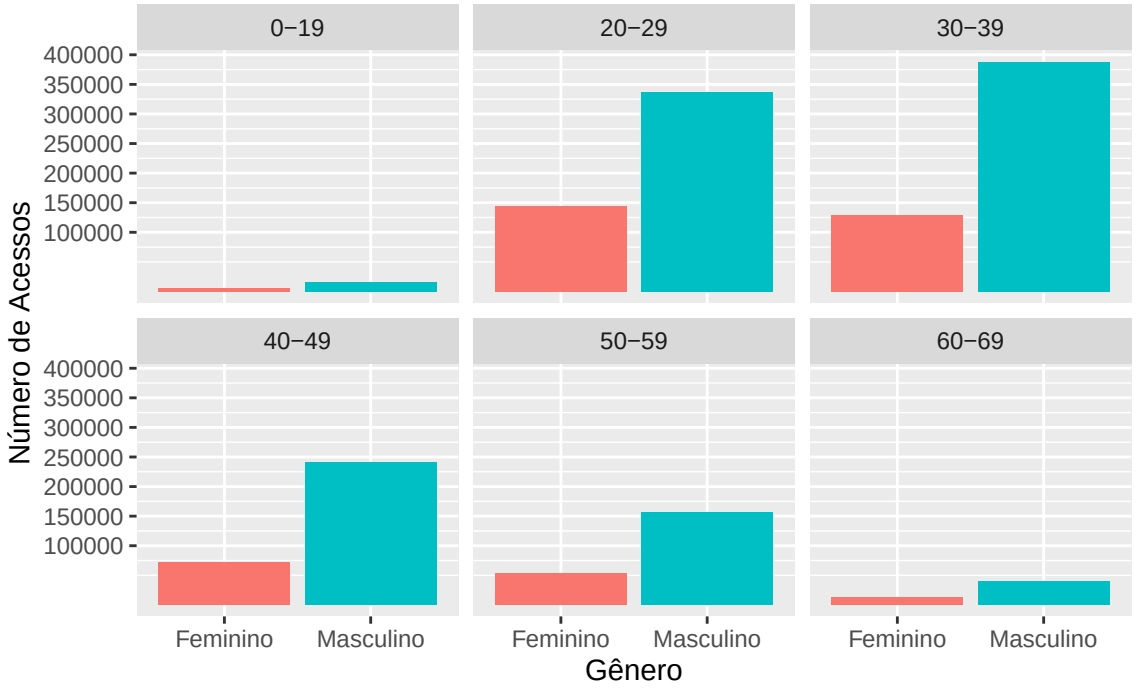
Pelo menos em Maio, quarta-feira foi o dia mais usado. Colocaria mais bicicletas na rua nestes dias.

O que mais podemos saber sobre os usuários?

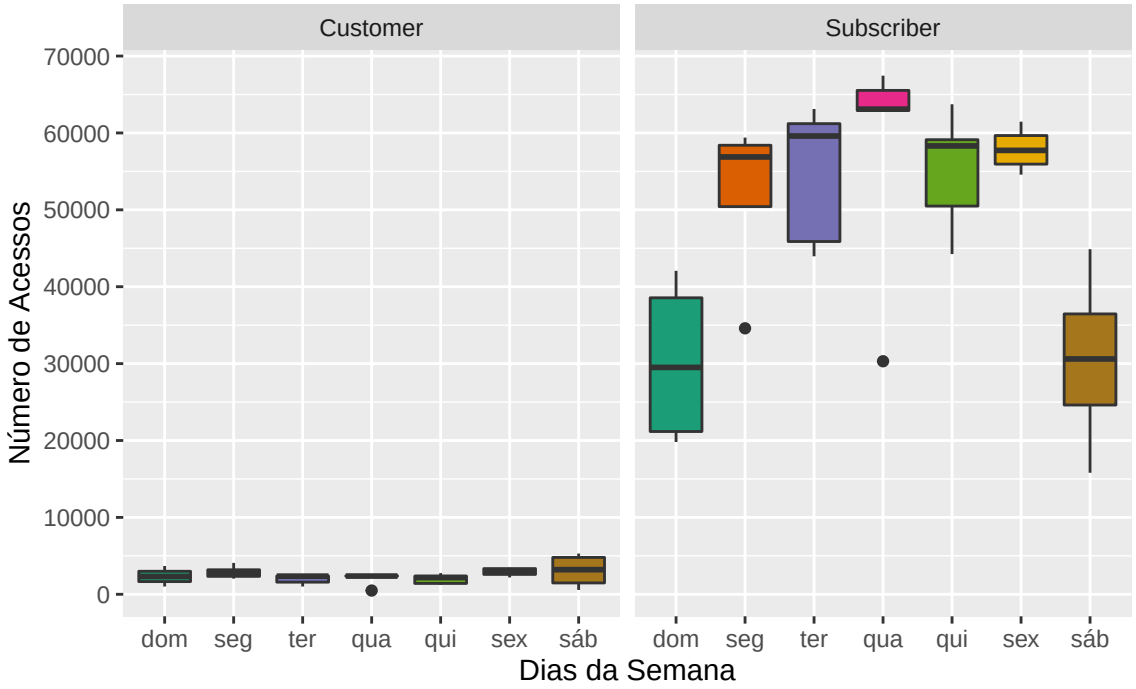
Independente do genêro, quarta-feira é o dia mais usado



Independente da faixa etária, os homens foram os que mais usaram o serviço



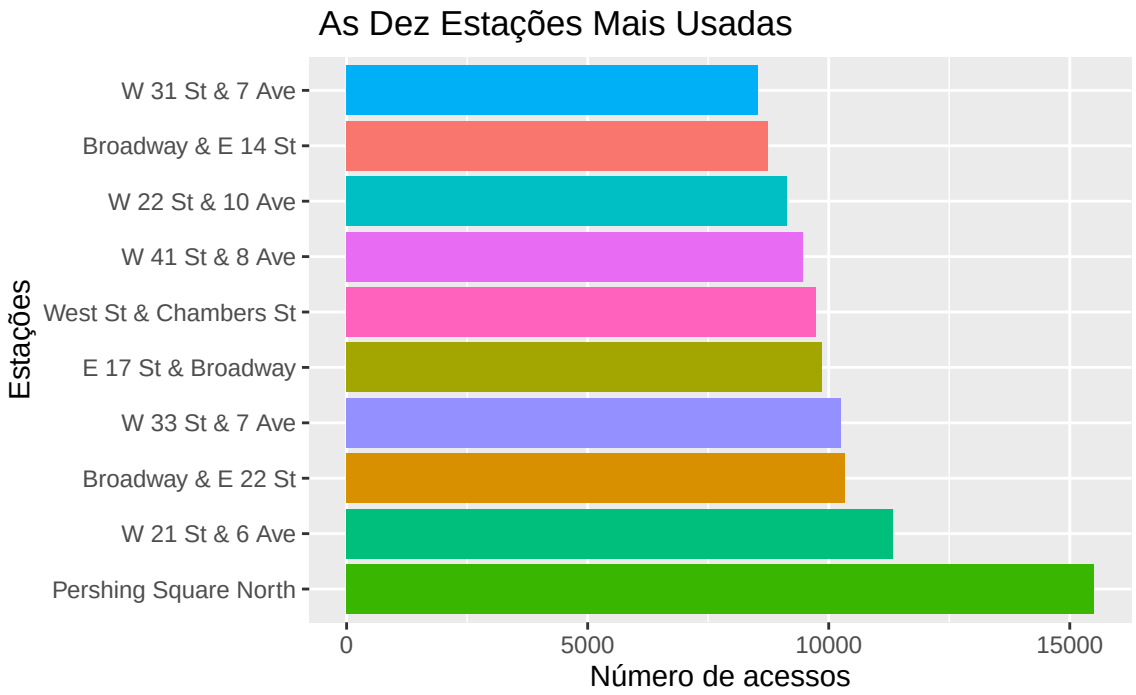
`Asssinantes`(*Subscriber*) além de utilizarem as bicicletas mais vezes que `clientes avulsos`(*Customer*), também utilizam mais nas quartas-feiras.



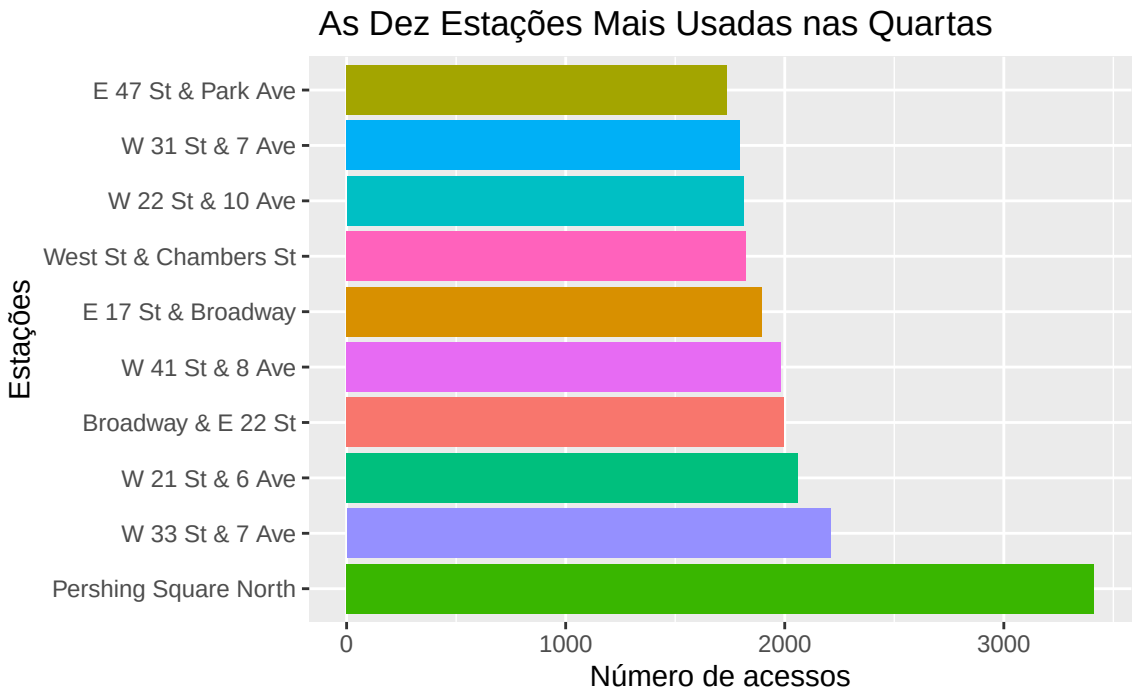
Curiosidade: Quando sou responsavél técnico em alguma drogaria (Sou farmacêutico de formação), é nítido a importância de um clinte fiel, que prefere consumir em sua drograria. Entretanto, este dataset deixa bem claro a diferença e como isso impacta os negocios e este artigo facilita o entendimento.

As Estações.

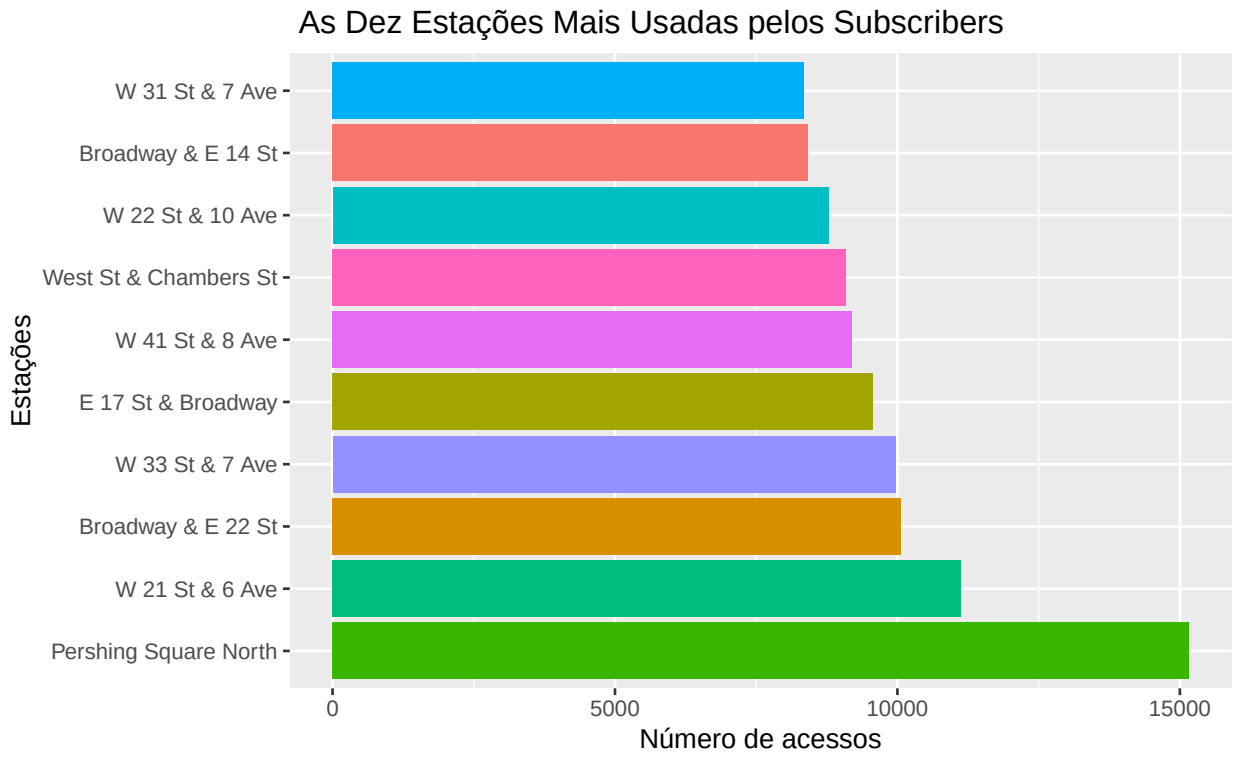
Agora iremos avaliar quais são as estações que mais tiveram acesso dos usuário em todo mês de Maio.

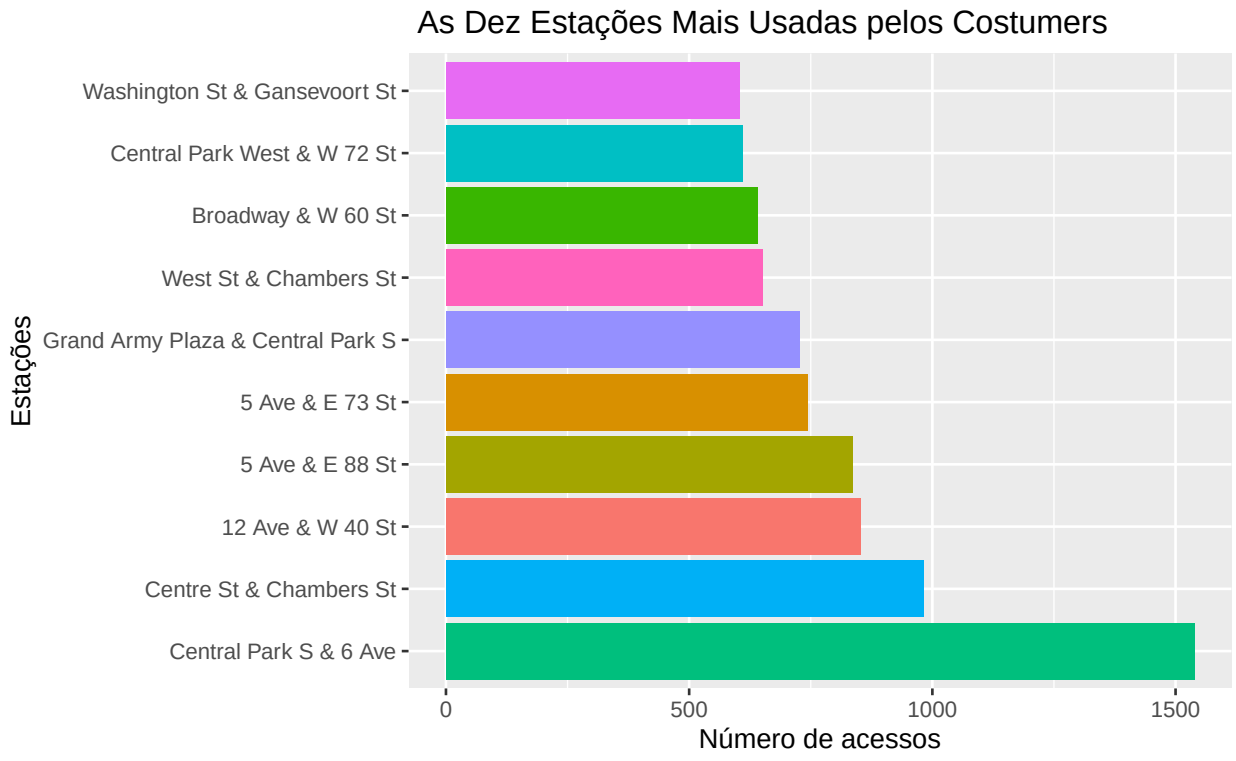


Qual a mais estação mais utilizada na quarta?



Há variação pelo tipo de usuário?





Vemos por essa última comparação que os consumidores avulsos tem utilizam estaçãos diferentes que os assinantes. Considerano que a mais uada pelos avulsos foi a do Central Park, eu imagino que sejá para passear por ele, apenas um achismos.

De quaquer forma, talvez seja valiada uma investigação mais profunda. E a partir do resultado, isto pode ser utilizado pelo marketing para alocação de propagrandas nestas estações, específicas para angariar mais inscrições ao programa.

Conclusão

Tivemos uma visão geral qual é o perfil e consumidores dos usuários de biclicletas alugadas em Nova Iorque e possiveis insights que podem ajudar a equipe de negocios a tomar decisões melhores.