

Convertendo Passageiros Casuais em Membros Anuais

Um caso de estudo em um sistema de compartilhamento de bicicletas

Por Bruno da Costa Calegari

Julho de 2022

Sumário

Objetivo do Projeto	3
Introdução	3
Cyclistic.....	3
Problema de negócio	4
1ª Etapa: Perguntar	5
2ª Etapa: Preparar	6
3ª Etapa: Processar	7
4ª Etapa: Analisar	11
5ª Etapa: Compartilhar	20
6ª Etapa: Agir	22

Objetivo do Projeto

Introdução

Neste estudo de caso proposto pelo Google vamos fazer uma análise sobre uma empresa fictícia chamada Cyclistic sobre um programa de compartilhamento de bicicletas, para responder os problemas de negócios propostos vamos fazer uso das seis etapas da análise de dados: perguntar, preparar, processar, analisar, compartilhar e agir.

Cyclistic

Em 2016, a Cyclistic lançou um programa bem-sucedido de compartilhamento de bicicletas. Desde então, o programa cresceu para uma frota de 5.824 bicicletas que são georastreadas e bloqueadas em uma rede de 692 estações em Chicago. As bicicletas podem ser desbloqueadas a partir de uma estação e devolvidas a qualquer outra estação da companhia a qualquer momento.

Até agora, a estratégia de marketing da Cyclistic era bem generalista, tentando um apelo a amplos segmentos de consumidores. Uma abordagem que ajudou a tornar isso possível foi a flexibilidade de seus planos de preços: passes de viagem única, passes de dia inteiro, e associações anuais. Os clientes que compram passes de viagem única ou de dia inteiro são chamados de passageiros casuais. Clientes que compram assinaturas anuais são membros do programa.

Os analistas financeiros da Cyclistic concluíram que os membros anuais são muito mais lucrativos do que os passageiros casuais. Apesar de flexibilidade de preços ajudar a Cyclistic a atrair mais clientes, acredita-se que maximizar o número de membros anuais pode ser a chave para um crescimento futuro. Em vez de criar uma campanha de marketing voltada para novos clientes, acredita-se que há uma boa chance de converter passageiros casuais em membros. Observa-se que os ciclistas casuais já estão cientes do programa de assinatura anual e escolheram Cyclistic para suas necessidades de mobilidade.

Problema de negócio

Com um objetivo claro de criar estratégias de marketing destinadas a converter passageiros casuais em membros anuais estabelecido, foram geradas três perguntas principais para guiar o futuro programa de marketing:

1. Como os membros anuais e os ciclistas casuais usam as bicicletas Cyclistic de forma diferente?
2. Por que os passageiros casuais comprariam assinaturas anuais da Cyclistic?
3. Como a Cyclistic pode usar a mídia digital para influenciar os ciclistas casuais a se tornarem membros?

Essas perguntas ajudarão a equipe de analistas de marketing a entender melhor como os membros anuais e os passageiros casuais usam as bicicletas, por que pilotos casuais comprariam uma assinatura anual e como a mídia digital poderia afetar suas táticas de marketing. Para responder essas perguntas efetivamente, vamos analisar os dados históricos das viagens de bicicletas de 2021 para identificar tendências e gerar insights que auxiliarão nas tomadas de decisões.



1ª Etapa: Perguntar

A primeira fase da análise de dados consiste em definir o problema a ser resolvido, no caso da companhia Cyclistic, temos o objetivo claro de converter passageiros casuais em membros anuais a fim de possibilitar um futuro crescimento.

Para guiar a nossa análise podemos formular questões usando o método SMART, esse método consiste em fazer perguntas:

- Específicas
- Mensuráveis
- Orientadas para a ação
- Relevantes
- Com duração definida.

Além do uso do método, deve se fazer perguntas abertas as partes interessadas, evitando perguntas sugestivas, fechadas ou vagas. Assim podemos entender melhor todo o contexto e ficar totalmente alinhado com as expectativas geradas.

Portanto, com a ajuda do método SMART, formulamos três perguntas essenciais para ser a alçada da nossa análise:

1. Como ciclistas casuais e membros anuais usam o programa de bicicleta compartilhada?
2. Como fazer com que ciclistas casuais se tornem membros anuais?
3. Como a Cyclistic pode usar a mídia digital para influenciar os ciclistas casuais a se tornarem membros?

A partir desses insights, a equipe de marketing criará uma nova estratégia converter passageiros casuais em membros anuais. Portanto nossa análise deve ser apoiada com insights de dados convincentes e visualizações de dados objetivas e relevantes.

Como já temos nossa tarefa de negócio bem definida e alinhada com as partes interessadas, podemos prosseguir para a próxima fase da análise.



2ª Etapa: Preparar

Com o nosso problema de negócio bem definido, vamos preparar os dados para a exploração. Iremos usar dados históricos de 2021 de viagens do Cyclistic para analisar e identificar tendências.

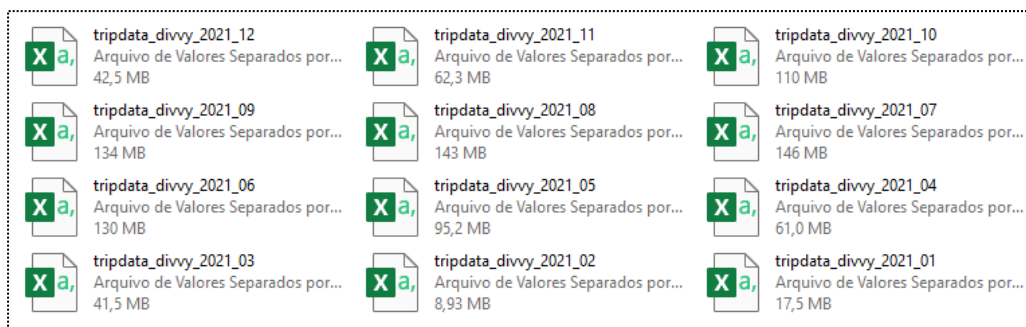
Como a Cyclistic é uma companhia fictícia criada apenas para esse estudo de caso, vamos usar dados fornecidos pela Motivate International Inc. sob esta [licença](#). Esses são dados públicos e apropriados para responder as tarefas de negócio, porém possuem algumas limitações por conta da privacidade de dados, nos proibindo de fazer uso informações de identificação pessoal dos passageiros.

Como já temos uma fonte de dados, vamos usar o método ROAAC para identificar se é uma boa fonte.

- R - Real
- O - Original
- A - Abrangente
- A - Atual
- C – Citado

Podemos ter certeza que estamos obtendo informações reais, originais, abrangentes (apenas com algumas limitações por conta da anonimização dos dados), atuais e bem citadas, já que estamos usando informações fornecidas por uma empresa real, logo podemos ficar seguros em relação a confiabilidade desses dados.

Vamos fazer o download desses dados dessa [fonte](#), armazenar, organizar e renomear os arquivos .csv com uma nomenclatura definida: ‘nomearquivo_aaaa_mm’.





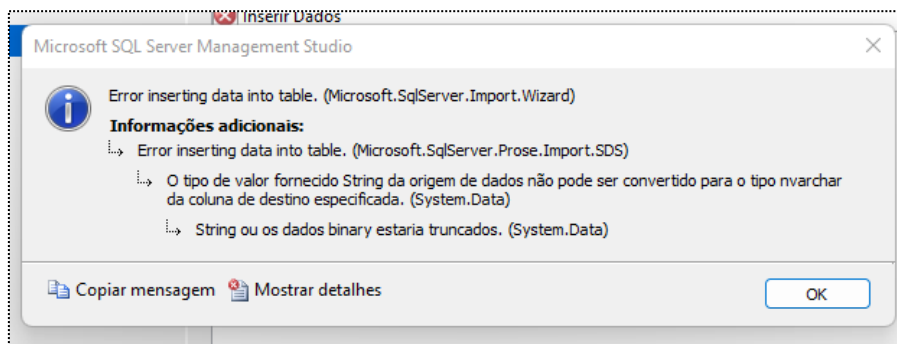
3ª Etapa: Processar

Nessa etapa de processamento, vamos fazer a limpeza dos dados para nos livrar de quaisquer erros, imprecisões ou possíveis inconsistências.

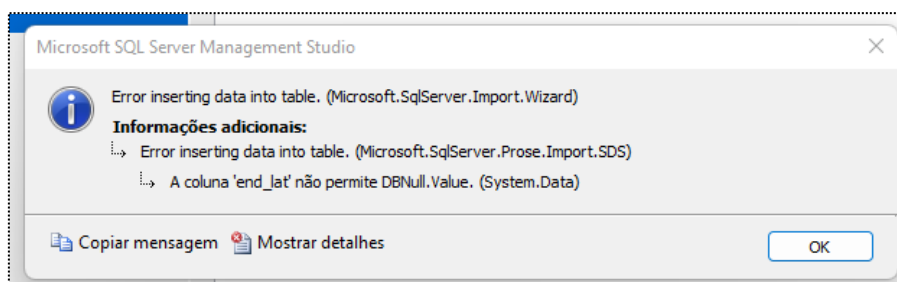
Após criar um banco de dados com o nome divvy no Microsoft SQL Server, vamos carregar todos os arquivos .csvs de 2021.

Ao tentar importar os arquivos obtemos dois tipos de erro:

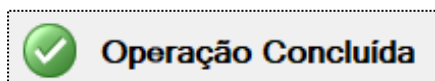
1.



2.



Para consertar esses dois erros vamos permitir valores nulos nas colunas que necessitam e usar o tipo de dado nvarchar(MAX):



Operação Concluída

Nome da Coluna	Tipo de Dados	Chave Primária	<input type="checkbox"/> Permitir Nulos
ride_id	nvarchar(MAX)	<input type="checkbox"/>	<input type="checkbox"/>
rideable_type	nvarchar(MAX)	<input type="checkbox"/>	<input type="checkbox"/>
started_at	datetime2	<input type="checkbox"/>	<input type="checkbox"/>
ended_at	datetime2	<input type="checkbox"/>	<input type="checkbox"/>
start_station_name	nvarchar(MAX)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
start_station_id	nvarchar(MAX)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
end_station_name	nvarchar(MAX)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
end_station_id	nvarchar(MAX)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
start_lat	float	<input type="checkbox"/>	<input type="checkbox"/>
start_lng	float	<input type="checkbox"/>	<input type="checkbox"/>
end_lat	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
end_lng	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
member_casual	nvarchar(MAX)	<input type="checkbox"/>	<input type="checkbox"/>

Com todas as tabelas já importadas para o nosso banco de dados, vamos criar uma nova tabela contendo todas as outras tabelas, essa vai ser a nossa tabela anual, contendo todos os meses de 2021:

```
-- CRIAR TABELA PARA ARMAZENAR TODOS OS DADOS DE 2021

CREATE TABLE dbo.tripdata_divvy_2021

(
    ride_id NVARCHAR(MAX),
    rideable_type NVARCHAR(MAX),
    started_at DATETIME2,
    ended_at DATETIME2,
    start_station_name NVARCHAR(MAX),
    start_station_id NVARCHAR(MAX),
    end_station_name NVARCHAR(MAX),
    end_station_id NVARCHAR(MAX),
    start_lat FLOAT,
    start_lng FLOAT,
    end_lat FLOAT,
    end_lng FLOAT,
    member_casual NVARCHAR(MAX),
    ride_length_s FLOAT,
    week_day FLOAT,
    y_month NVARCHAR(MAX),
)
```

Agora vamos inserir os dados na nossa tabela com o seguinte código:

```
-- INSERIR DADOS NA TABELA CRIADA

INSERT INTO dbo.tripdata_divvy_2021
SELECT *,
    DATEDIFF(second, started_at, ended_at), -- Calcular tempo do passeio
    DATEPART(weekday, started_at), -- Calcular dia da semana, 1=Domingo, 7=Sábado
    'Jan' -- Mudar o mês a cada tabela
FROM dbo.tripdata_divvy_2021_01 -- Repetir para cada tabela
```

Esse código irá inserir todos os dados referentes a tabela que especificamos e adicionar valores as três colunas extras que criamos na nossa tabela anual:

ride_length_s → Tempo do passeio em segundos.

week_day → Dia da semana, onde 1 equivale a Domingo e 7 a Sábado.

y_month → Mês de referência, usamos isso para separar os dados importados das tabelas mensais.

Vamos repetir esse código para cada tabela mensal, sempre mudando o mês e a tabela mensal.

Com a nossa tabela anual criada, carregada e classificada com todos os dados provenientes das tabelas mensais podemos prosseguir para a limpeza desses dados.

Sabemos que duas colunas devem possuir valores específicos:

'rideable_type' = 'electric_bike', 'classic_bike' e 'docked_bike'

'member_casual' = 'casual' e 'member'

Então vamos verificar se não existem valores digitados incorretamente:

```
-- Verificar se existem valores digitados incorretamente.  
  
SELECT DISTINCT  
    rideable_type  
FROM  
    dbo.tripdata_divvy_2021  
SELECT DISTINCT  
    member_casual  
FROM  
    dbo.tripdata_divvy_2021
```

Ao obter os resultados, podemos ver que não existem valores digitados incorretamente.

1.		rideable_type
	1	electric_bike
	2	classic_bike
	3	docked_bike

2.		member_casual
	1	member
	2	casual

Então vamos remover os possíveis espaços em branco em todas as colunas que contém strings:

```
-- Remover espaços em branco de todas as colunas que contém strings  
  
UPDATE  
    dbo.tripdata_divvy_2021  
SET  
    ride_id = TRIM(ride_id),  
    start_station_name = TRIM(start_station_name),  
    start_station_id = TRIM(start_station_id),  
    end_station_name = TRIM(end_station_name),  
    end_station_id = TRIM(end_station_id);
```

Como vimos na importação de dados para nosso banco de dados, possuímos seis colunas com dados nulos, são colunas referentes sobre aonde começou e terminou cada viagem, como não queremos que esses dados impactem em nossa análise, vamos finalizar nossa limpeza deletando todas as linhas que contenham dados nulos nessas colunas:

```
-- Deletar todas as linhas que contenham valores nulos  
  
DELETE FROM dbo.tripdata_divvy_2021  
WHERE start_station_name IS NULL OR start_station_id IS NULL OR end_station_name IS  
NULL OR end_station_id IS NULL OR end_lat IS NULL OR end_lng IS NULL;
```

A coluna 'ride_id' contém valores únicos, então vamos checar se não existem valores duplicados nela:

```
-- Verificar se existem valores duplicados

SELECT TOP 10
    ride_id,
    COUNT(ride_id) AS cont
FROM
    dbo.tripdata_divvy_2021
GROUP BY
    ride_id
HAVING
    COUNT(ride_id) > 1;
```

Logo, podemos ver que não existem valores duplicados:

```
(0 linhas afetadas)

Horário de conclusão: 2022-07-20T19:57:29.8746418-03:00
```

Agora, vamos verificar nossa coluna 'ride_length_s' para ver se existem valores incorretos, como tempos de viagens negativos ou muito baixos, lembrando que o tempo de viagem está em segundos. Vamos limitar nosso escopo de análise apenas para viagens maiores que um minuto, logo vamos deletar todas as linhas que possuam valores menores ou iguais a 60 segundos:

```
-- Deletar valores inferiores ou iguais a 60 segundos na coluna tempo de viagem

DELETE FROM dbo.tripdata_divvy_2021
WHERE
    ride_length_s <= 60
```

R:

```
(60011 linhas afetadas)

Horário de conclusão: 2022-07-20T20:13:27.9721906-03:00
```



4ª Etapa: Analisar

Com os dados limpos, podemos olhar e explorar nosso conjunto de dados para fazer conexões e identificar tendências, como já agregamos nossos dados quando criamos a nossa tabela anual e classificamos por mês, vamos dar uma olhada geral nos valores médio, máximo e mínimo do tempo de viagem.

```
SELECT
  AVG(ride_length_s) AS media,
  MAX(ride_length_s) AS maximo,
  MIN(ride_length_s) AS minimo
FROM
  dbo.tripdata_divvy_2021
```

R:

	media	maximo	minimo
1	1325,66708698712	3356649	61

Podemos ver que a média, o máximo e o mínimo anual são de:

Média Anual	Máximo Anual	Mínimo Anual
+/- Vinte e dois minutos e seis segundos	Oito horas, vinte e quatro minutos e nove segundos.	Um minuto e um segundo

Em qual dia da semana ocorre a maior frequência de passeios:

```
-- Moda dos dias da semana
SELECT
  week_day,
  COUNT(week_day) AS frq_week_day
FROM
  dbo.tripdata_divvy_2021
GROUP BY
  week_day
ORDER BY
  frq_week_day DESC;
```

R:

	week_day	frq_week_day
1	7	814736
2	1	705337
3	6	647322
4	4	607914
5	3	595231
6	5	590065
7	2	567686

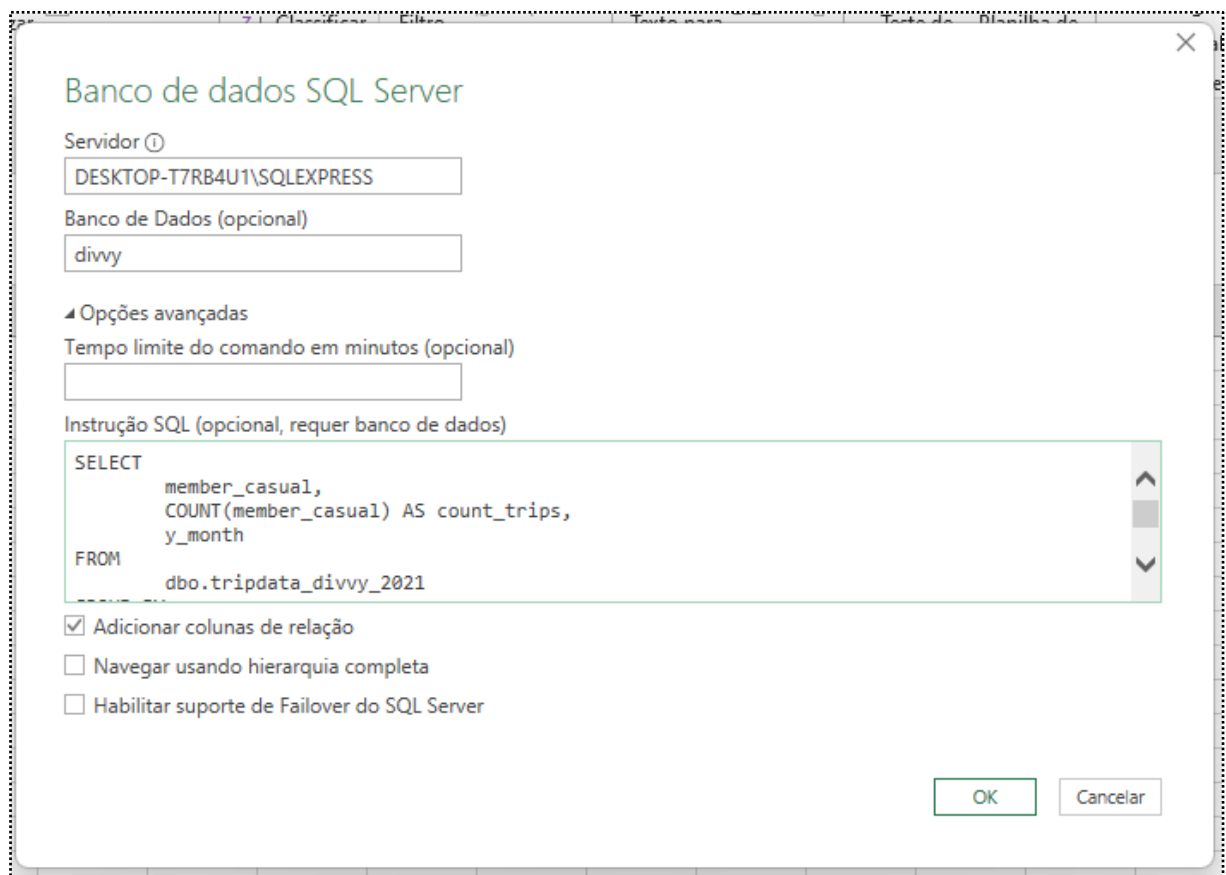
Os finais de semana são os dias de maior frequência, Sábado e Domingo respectivamente.

Agora que já vimos o básico do nosso conjunto de dados, vamos realizar análises mais específicas para descobrir qual a diferença no uso do programa entre ciclistas casuais e membros anuais do programa da Cyclistic.

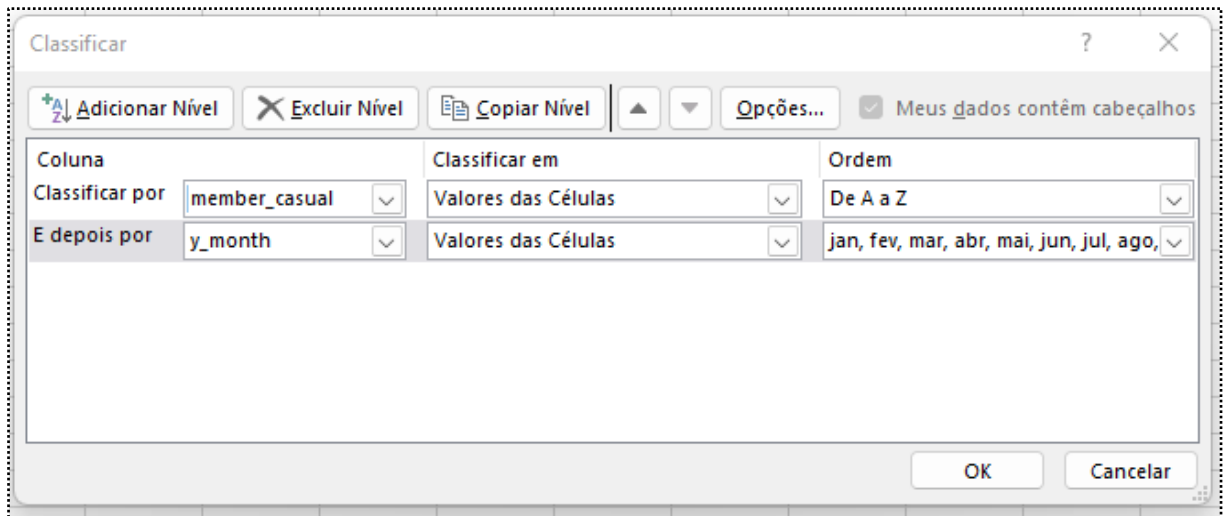
Vamos começar pelas viagens por mês entre ciclistas casuais e membros anuais:

```
-- Viagens por mês entre ciclistas casuais e membros anuais
SELECT
    member_casual,
    COUNT(member_casual) AS count_trips,
    y_month
FROM
    dbo.tripdata_divvy_2021
GROUP BY
    member_casual, y_month;
```

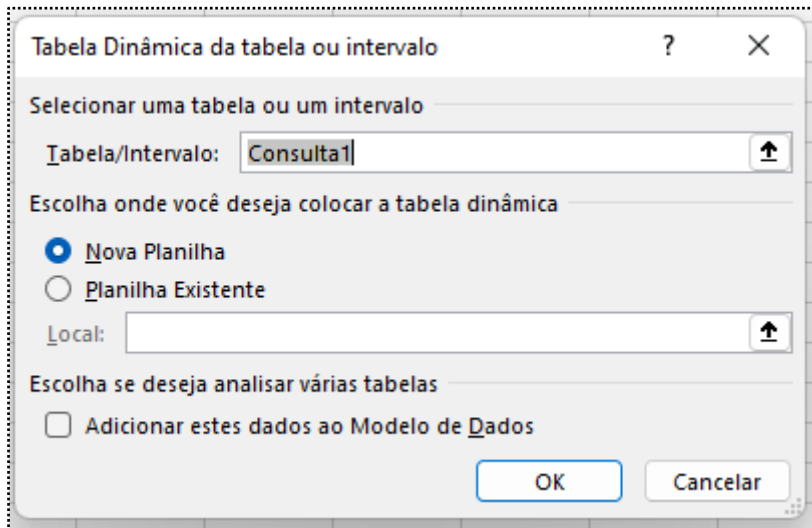
Ao invés de realizar a consulta no SQL, vamos realizar direto no excel, para classificar e gerar uma visualização:



Classificando os dados pelo tipo de usuário e mês:



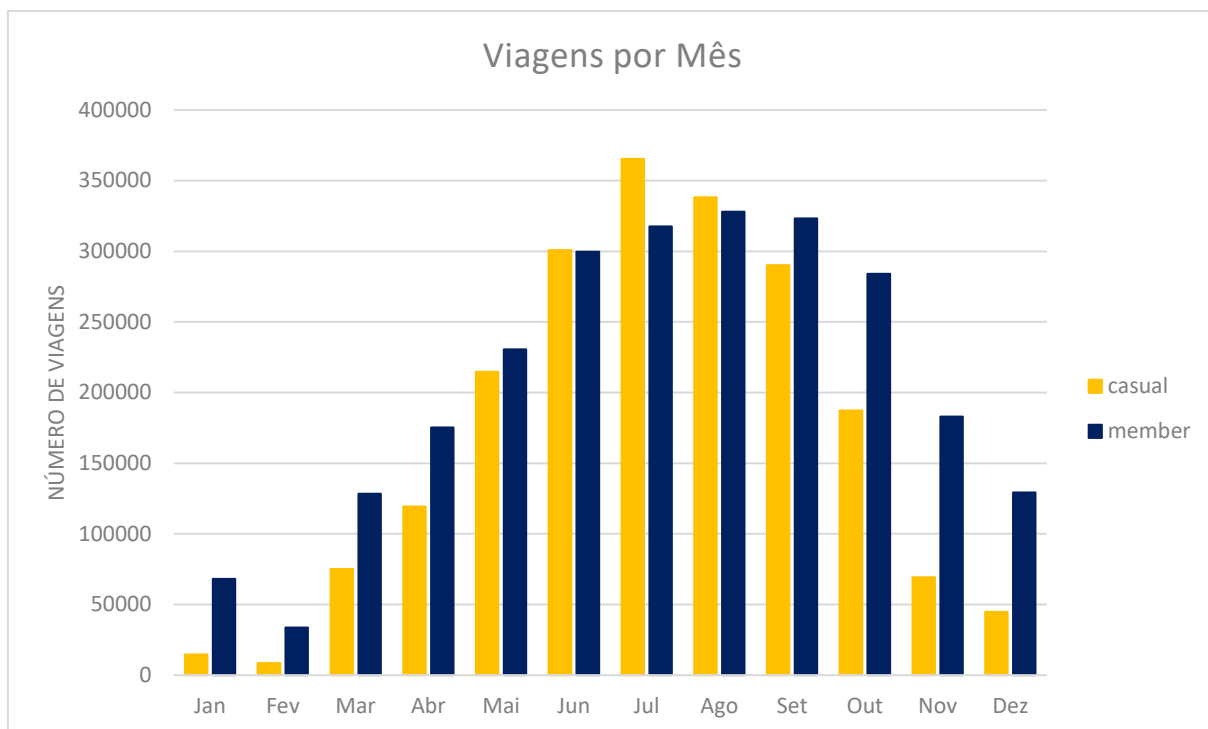
Gerando uma tabela dinâmica para analisar melhor a consulta:



Organizando tabela dinâmica:

Soma de count_trips			
Rótulos de Coluna			
Rótulos de Linha	casual	member	Total Geral
Jan	14582	68030	82612
Fev	8506	33787	42293
Mar	75050	128333	203383
Abr	119349	175248	294597
Mai	214625	230470	445095
Jun	300735	299694	600429
Jul	365457	317633	683090
Ago	338095	327900	665995
Set	290043	323240	613283
Out	187302	284080	471382
Nov	69262	182892	252154
Dez	44682	129296	173978
Total Geral	2027688	2500603	4528291

Gerando o gráfico:



Usando de base as estações no hemisfério norte, para os usuários casuais podemos visualizar um crescimento na quantidade de viagens na entrada da primavera em março, o crescimento fica ainda mais acentuado quando o verão entra em junho, tendo seu pico máximo em julho. As viagens começam a cair na chegada do outono em setembro e com a chegada do inverno em dezembro, tendo seus picos máximos entre janeiro e dezembro.

Já para os membros anuais também podemos ver essa diferença de crescimento no número das viagens com a chegada da primavera e do verão, e sua queda com a chegada do outono e posteriormente o inverno, tendo suas piores quedas entre janeiro e fevereiro, porém de uma forma mais leve.

Conseguimos concluir que as estações do ano mais quentes como primavera e principalmente o verão, que também é associado a um mês de férias, atraem mais usuários casuais.

Para confirmar nossa análise, vamos executar duas consultas para obter a contagem de viagens dos membros anuais e dos usuários casuais por estações do ano:



1.

```
-- Contagem de viagens por estações do ano dos membros anuais
SELECT
  (SELECT
    (COUNT(CAST(started_at AS DATE)))
  FROM
    dbo.tripdata_divvy_2021
  WHERE
    member_casual = 'member' AND
    (started_at BETWEEN '2021-01-01' AND '2021-03-20'
    OR started_at BETWEEN '2021-12-22' AND '2021-12-31')) AS inverno_member,
  (SELECT(COUNT(CAST(started_at AS DATE)))
  FROM
    dbo.tripdata_divvy_2021
  WHERE
    member_casual = 'member' AND
    started_at BETWEEN '2021-03-20' AND '2021-06-21') AS primavera_member,
  (SELECT(COUNT(CAST(started_at AS DATE)))
  FROM
    dbo.tripdata_divvy_2021
  WHERE
    member_casual = 'member' AND
    started_at BETWEEN '2021-06-21' AND '2021-09-23') AS verao_member,
  (SELECT(COUNT(CAST(started_at AS DATE)))
  FROM
    dbo.tripdata_divvy_2021
  WHERE
    member_casual = 'member' AND
    started_at BETWEEN '2021-09-23' AND '2021-12-22') AS outono_member;
```

2.

```
-- Contagem de viagens por estações do ano dos usuários casuais
SELECT
  (SELECT
    (COUNT(CAST(started_at AS DATE)))
  FROM
    dbo.tripdata_divvy_2021
  WHERE
    member_casual = 'casual' AND
    (started_at BETWEEN '2021-01-01' AND '2021-03-20'
    OR started_at BETWEEN '2021-12-22' AND '2021-12-31')) AS inverno_casual,
  (SELECT(COUNT(CAST(started_at AS DATE)))
  FROM
    dbo.tripdata_divvy_2021
  WHERE
    member_casual = 'casual' AND
    started_at BETWEEN '2021-03-20' AND '2021-06-21') AS primavera_casual,
  (SELECT(COUNT(CAST(started_at AS DATE)))
  FROM
    dbo.tripdata_divvy_2021
  WHERE
    member_casual = 'casual' AND
    started_at BETWEEN '2021-06-21' AND '2021-09-23') AS verao_casual,
  (SELECT(COUNT(CAST(started_at AS DATE)))
  FROM
    dbo.tripdata_divvy_2021
  WHERE
    member_casual = 'casual' AND
    started_at BETWEEN '2021-09-23' AND '2021-12-22') AS outono_casual;
```

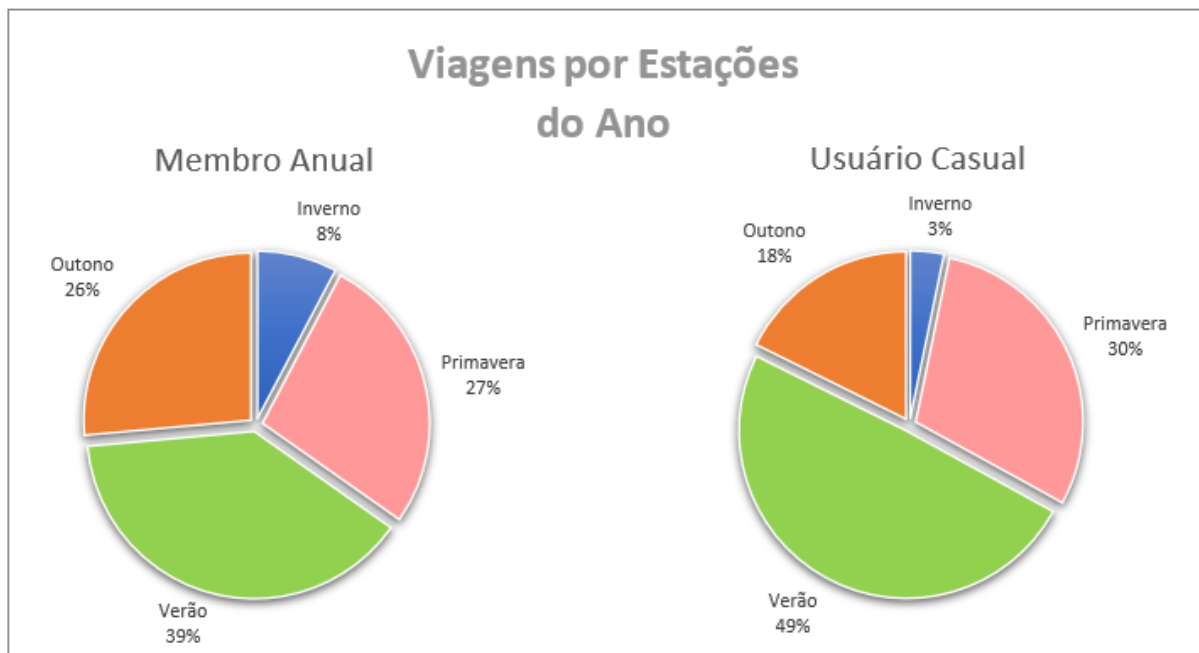

Realizando as consultas direto no excel, fizemos duas conexões:

	Consulta2
	Apenas conexão.
	Consulta3
	Apenas conexão.

Agora carregamos os dados para uma planilha, para formata-los em uma tabela criada para melhor visualização:

					Estação do Ano	Membro	Casual
					Inverno	192865	65680
					Primavera	675161	602909
					Verão	969578	999674
					Outono	659968	357830
inverno_memb	primavera_memb	verao_memb	outono_memb				
192865	675161	969578	659968				
inverno_casual	primavera_casual	verao_casual	outono_casual				
65680	602909	999674	357830				

Gerando dois gráficos de pizza, um para os membros e outro para os usuários casuais, fizemos um painel básico das viagens por estações do ano:



De modo geral, confirmamos que o inverno é a pior estação e o verão a melhor para as viagens, principalmente para os usuários casuais. Notamos também que para os membros anuais, no

outono e na primavera o número de viagens é semelhante, enquanto que para os usuários casuais as viagens no outono são bem menores do que na primavera.

Após analisarmos o número de viagens por mês e por estações do ano, vamos aproximar a nossa lupa e ver como essas viagens se distribuem durante a semana para os membros anuais e para os usuários casuais. Vamos executar nossa consulta SQL importando os dados para o excel:

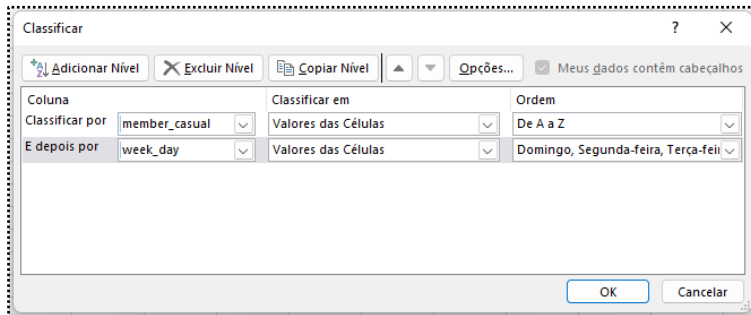
```
-- Qtd de viagens por dia da semana para membros e usuários casuais

SELECT
    member_casual,
    week_day,
    COUNT(week_day) AS trips
FROM
    dbo.tripdata_divvy_2021
GROUP BY
    member_casual, week_day
ORDER BY
    member_casual;
```

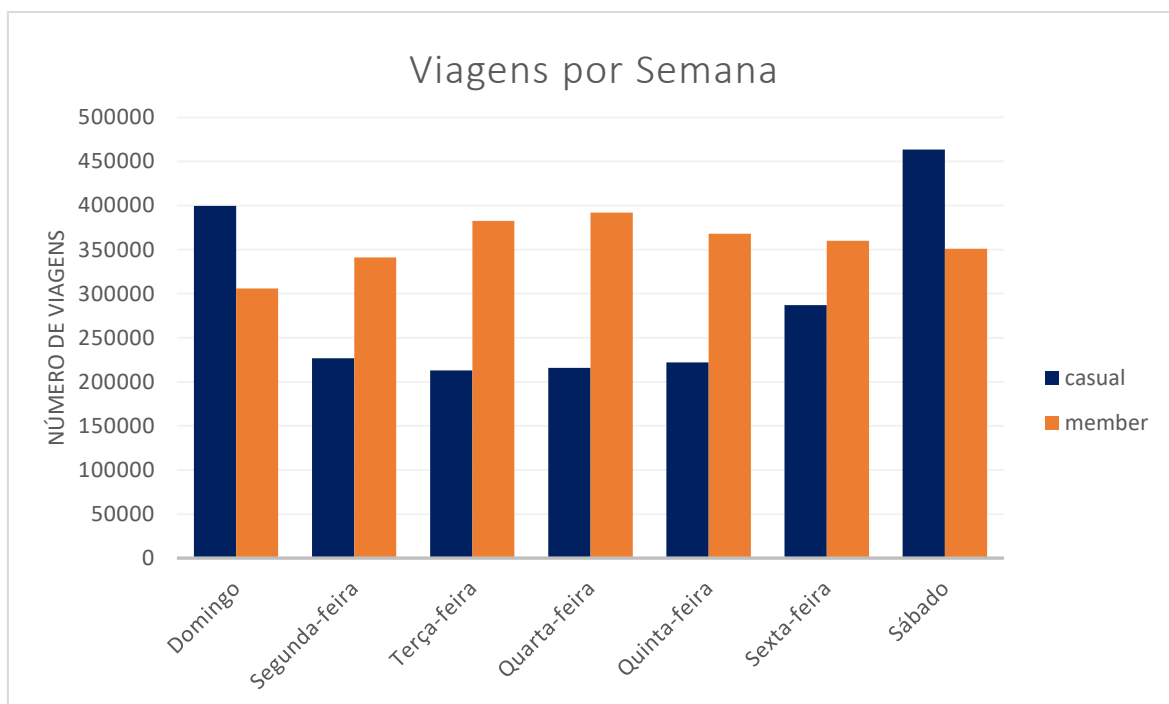
Vamos fazer o tratamento dos dados, convertendo a coluna 'week_day' para o formato "texto" e fazendo a substituição dos valores, lembrando que 1 = Domingo e 7 = Sábado:



Após o tratamento dos dados, vamos carregar nossa consulta em uma nova planilha e fazer a classificação:



Com os dados tratados e classificados, vamos resumir em uma tabela dinâmica e gerar o nosso gráfico:



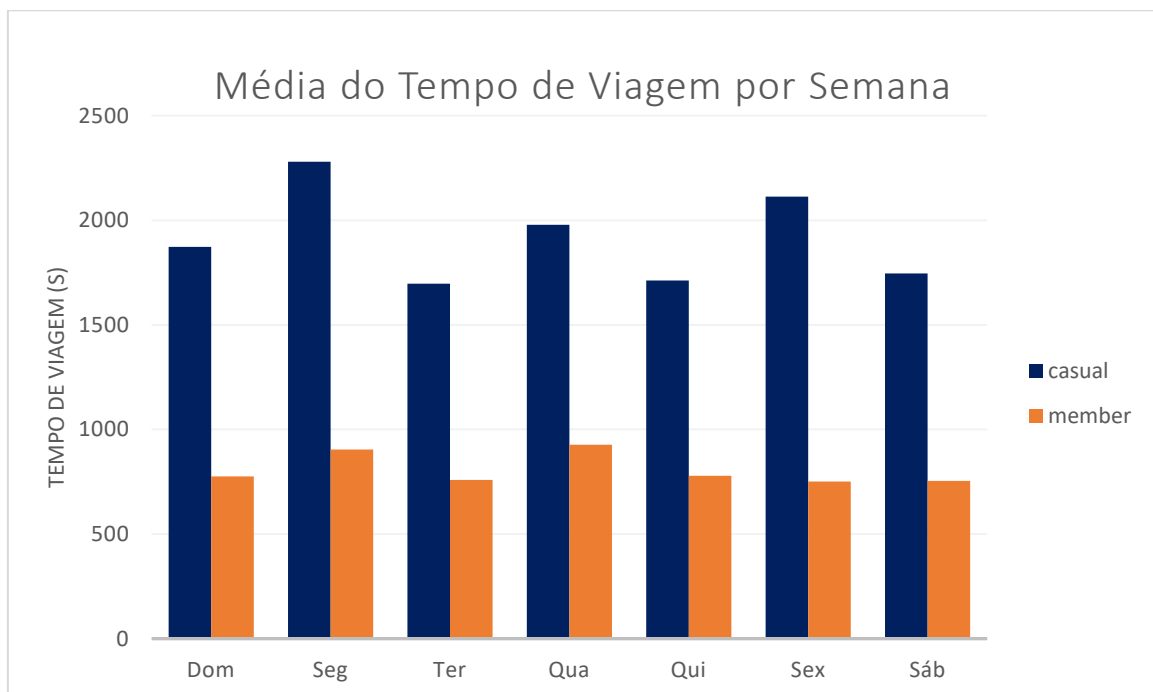
Enquanto membros anuais mantêm um número consistente de viagens durante a semana, usuários casuais fazem um maior número de viagens aos finais de semana.

Para confirmar nossas descobertas, vamos efetuar nossa última análise, sobre o tempo de viagem médio para membros anuais e usuários casuais por semana:

```
-- Média de tempo de viagem para usuários casuais e membros anuais por semana
```

```
SELECT
    member_casual,
    week_day,
    AVG(ride_length_s) AS avg_ride
FROM
    dbo.tripdata_divvy_2021
GROUP BY
    member_casual, week_day
ORDER BY
    member_casual
```

Após carregar nossa consulta no Excel, vamos tratar os dados da coluna 'week_day' como na consulta anterior, classificar e resumir em uma tabela dinâmica para gerar nosso gráfico:



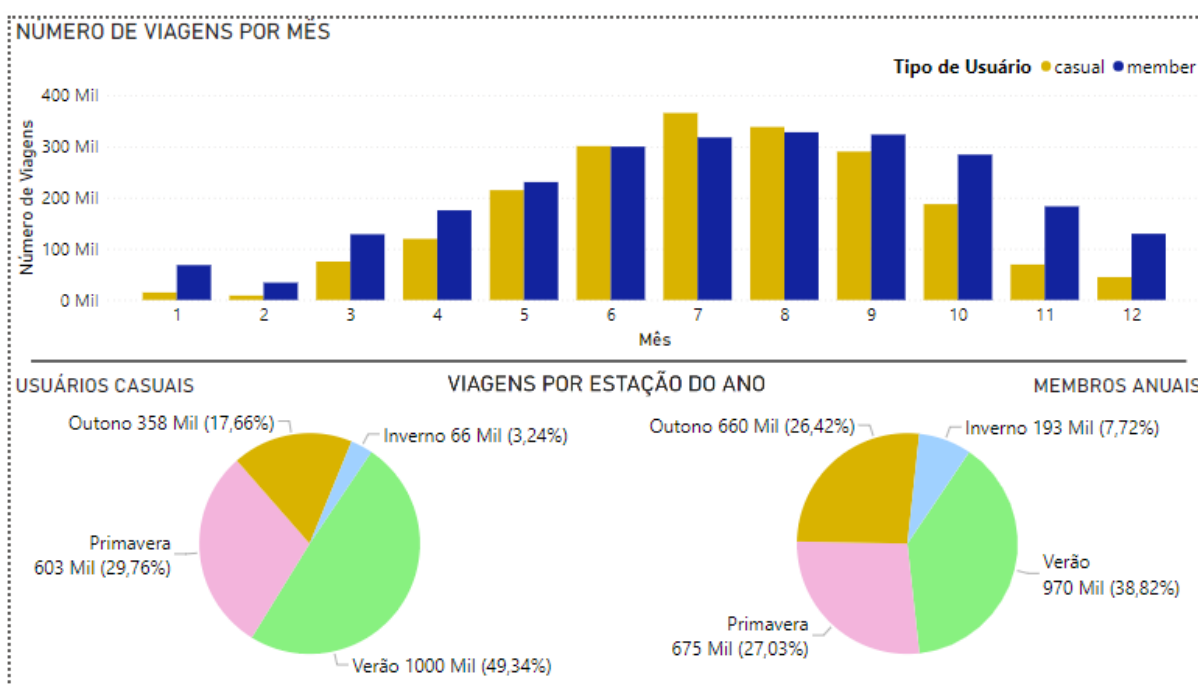
O tempo médio de corrida se mantém mais constante durante a semana para os membros, enquanto que para os usuários casuais tem uma mudança de valor mais alta entre cada dia da semana. Por fim, podemos concluir que os usuários casuais costumam fazer corridas mais longas do que os membros anuais.



5ª Etapa: Compartilhar

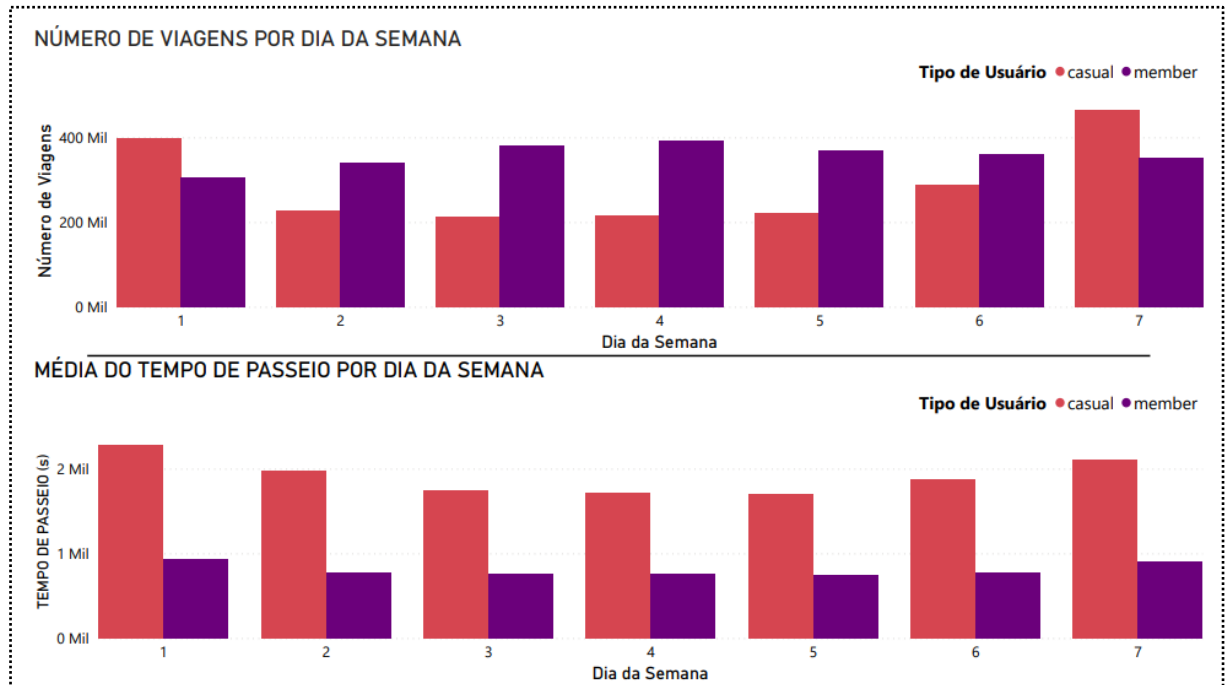
Após realizar a análise e ver o que os dados nos contam, vamos compartilhar nossas principais conclusões com visualizações polidas e objetivas. Com o Power BI vamos gerar dois painéis para suportar nossas principais conclusões:

1. Variação no número de viagens durante o ano e por estação para usuários casuais e membros anuais:



- ✓ O terceiro trimestre de 2021 é o melhor em número de viagens para ambos tipos de usuário.
- ✓ O número de viagens dos usuários casuais é mais afetado negativamente do que o dos membros anuais, porém também são mais afetados positivamente pelo verão em relação aos membros anuais.
- ✓ Podemos ver uma distribuição mais discrepante no número de viagens dos usuários casuais, que tem sua maior concentração no verão, enquanto que as viagens dos membros anuais, possuem uma distribuição mais uniforme entre o outono, primavera e verão.

2. Variação no número de viagens e média do tempo de passeio (em segundos) por dia da semana (1 = Domingo, 7 = Sábado) para usuários casuais e membros anuais:



- ✓ Enquanto membros anuais possuem seus números de viagens distribuídos de forma mais uniforme durante a semana, usuários casuais costumam fazer mais viagens aos finais de semana, podemos ver que o aumento começa a partir de sexta-feira.
- ✓ Usuários casuais costumam fazer viagens mais longas do que membros anuais.
- ✓ Os finais de semana são os dias com as viagens mais longas para os usuários casuais.



6ª Etapa: Agir

Na última etapa do processo de análise de dados, vamos fornecer nossas principais recomendações com base em nossas descobertas respondendo o nosso problema de negócio:

- Como fazer com que ciclistas casuais se tornem membros anuais?
 - I. Criar planos mais flexíveis, como assinaturas trimestrais, essas assinaturas poderão ser mais atraentes para usuários casuais, visto que o maior número de viagens se dá nos meses de verão.
 - II. Nos meses de verão, oferecer descontos na assinatura de planos anuais (ou trimestrais) visando atrair mais usuários casuais para se tornarem membros do programa, a campanha de marketing pode ser feita nos meses de inverno a primavera.
 - III. Criar um sistema de pontuação que contabilize o número de viagens e o tempo de cada passeio, os membros poderão trocar esses pontos por prêmios ou descontos. Membros anuais ou trimestrais que renovarem seus planos também ganharão pontos, com esse sistema de acumulação de pontos visamos fidelizar os membros obtidos.