media
.monks

Cauã Palugan Coelho

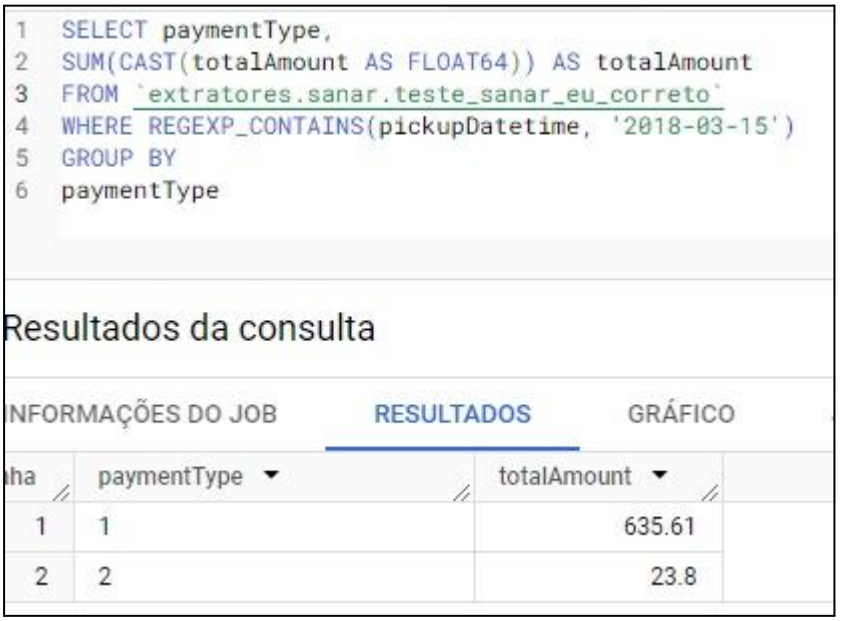
Relatório e Análise de Dados dos Casamentos Homoafetivos

Processo Seletivo Interno
Data Science

Projeto em SQL

Perguntas:

1) Qual foi a receita de cada tipo de pagamento no dia 15 de Março de 2018?
   R = **Cartão de Crédito** = $635,61
   Dinheiro = $23,80



```
1  SELECT paymentType,
2  SUM(CAST(totalAmount AS FLOAT64)) AS totalAmount
3  FROM `extratores.sanar.teste_sanar_eu_correto`
4  WHERE REGEXP_CONTAINS(pickupDatetime, '2018-03-15')
5  GROUP BY
6  paymentType
```

Resultados da consulta

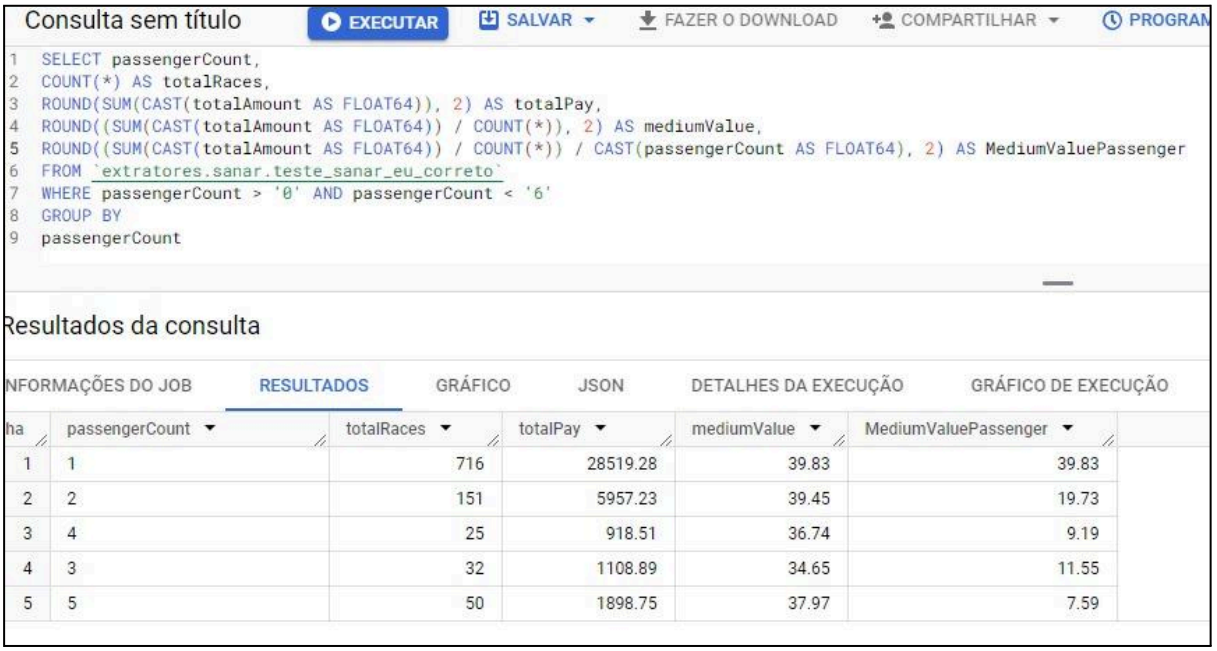
INFORMAÇÕES DO JOB | RESULTADOS | GRÁFICO

linha	paymentType	totalAmount
1	1	635.61
2	2	23.8

Explicação:

O que fiz nessa consulta foi, selecionar o Tipo de Pagamento, e o Valor Total pago por corrida, fiz um tratamento com o CAST transformando o Valor Total para Número com o FLOAT64.
Depois disso, com o WHERE e o REGEXP_CONTAINS, criei um filtro onde no campo pickupDatetime teria que conter '2018-03-15', que seria o dia 15 de Março de 2018 conforme solicitado.
Por fim só agrupei por tipo de pagamento as informações.

2) Considere que corridas de táxi válidas tenham de 1 a 5 passageiros. Qual a quantidade de corridas feitas com cada número de passageiros, faturamento médio de cada corrida e faturamento médio por passageiro?
R = Resposta na Imagem Abaixo



ha	passengerCount	totalRaces	totalPay	mediumValue	MediumValuePassenger
1	1	716	28519.28	39.83	39.83
2	2	151	5957.23	39.45	19.73
3	4	25	918.51	36.74	9.19
4	3	32	1108.89	34.65	11.55
5	5	50	1898.75	37.97	7.59

Explicação:
Comecei a consulta chamando o 'passengerCount', logo depois disso fiz um COUNT(*) e dei o nome de 'totalRaces' para sabermos qual o total de corridas foram feitas conforme o anunciado solicitou.
Depois disso fiz a soma do 'totalAmount' para vermos o valor total arrecadado nas corridas, para essa soma acontecer, fiz um tratamento nele com o CAST() para transformar ele em FLOAT64, para melhor visualização, usei o ROUND() para deixar apenas duas casas decimais no resultado final e renomeie ele para 'totalPay'.
Na próxima linha meu objetivo era coletar o Valor médio por corrida, para isso usei a mesma linha que tinha usado para fazer o 'totalPay', porém dividindo pelo número de corridas no caso foi a mesma linha utilizada para coletar o 'totalRaces', renomeei para 'mediumValue'.
Na linha seguinte gostaria de coletar o Valor médio por Passageiro em cada corrida, para isso utilizei a mesma linha utilizada no 'mediumValue', porém dividindo ela pelo 'passengerCount' também, para conseguir dividir ela pelo 'passengerCount' fiz um tratamento nele utilizando o CAST() para transformá-lo em FLOAT64 e depois disso renomeei essa linha para 'MediumValuePassenger'.
Lembrando que nesse SELECT usei o ROUND() na maioria das linhas para abreviar a quantidade de casas decimais.
Após tudo isso usando o WHERE fiz um filtro onde as únicas corridas que seriam selecionadas eram as que tivessem um número de passageiros maior que 0 e menor que 6.
Para finalizar gostaria de obter essas informações com base na quantidade de passageiros e por isso agrupei tudo por 'passengerCount'.

3) Qual a hora que mais começaram corridas?
R= A hora que mais se iniciaram corridas foi as 22:00Hrs, com 70 corridas.

```
1  SELECT EXTRACT(HOUR FROM TIMESTAMP(pickupDatetime)) AS hora,
2  COUNT(*) AS numero_corridas
3  FROM `extratores.sanar.teste_sanar_eu_correto`
4  GROUP BY
5  hora
6  ORDER BY
7  numero_corridas DESC
```

Resultados da consulta

INFORMAÇÕES DO JOB | RESULTADOS | GRÁFICO | JSON

nha	hora	numero_corridas
1	22	70
2	23	68
3	17	61

Explicação:

Comecei selecionando o campo pickupDateTime e fazendo um tratamento nele, estou utilizando o EXTRACT para extrair a hora do elemento que está na coluna, para isso faço um 'HOUR FROM TIMESTAMP' que transforma o campo em uma medida de tempo.
Após isso fiz um COUNT(*) e nomeei de 'numero_corridas' para termos como base quantas corridas foram feitas nessa mesma hora.
Depois disso só agrupei pela Hora que eu tinha extraído e ordenei pelo número de corridas para acharmos qual tinha sido a hora com mais corridas.

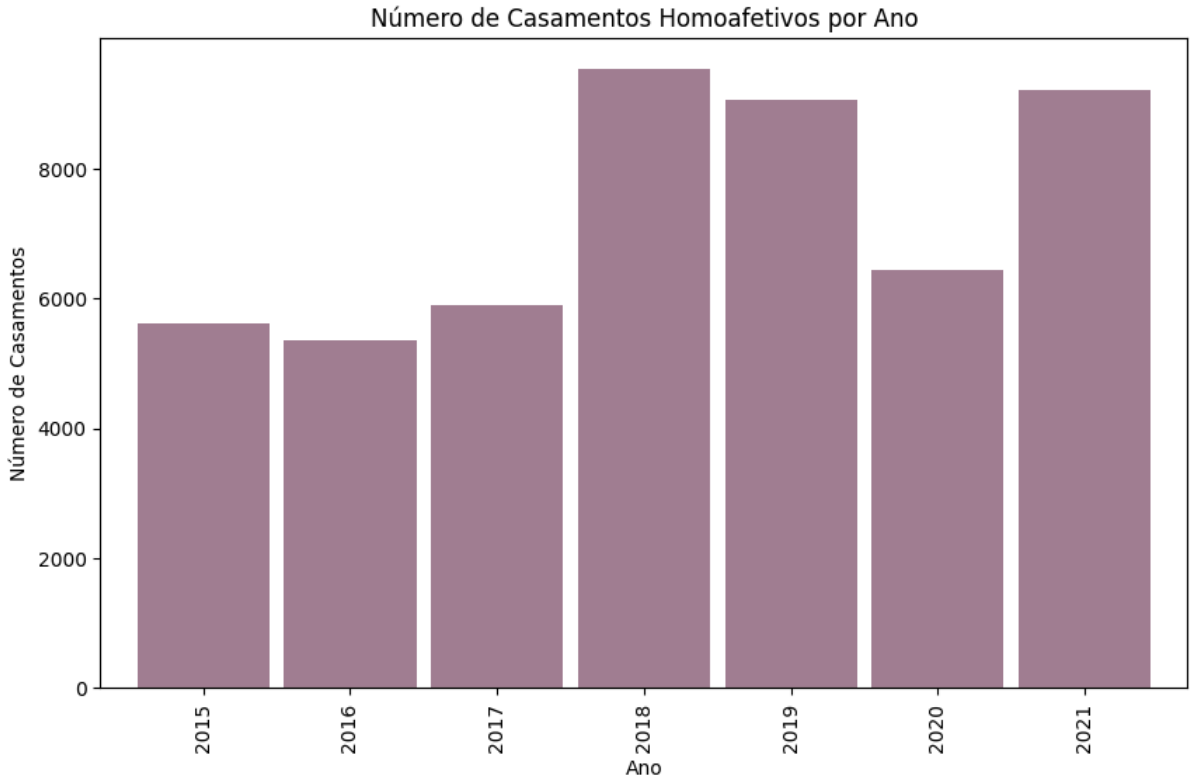
4) Considerando apenas as corridas que houveram pedágios (tolls) e que transportaram até 3 passageiros, qual a média do valor pago em pedágios por corrida?
R= A média do valor de pedágios pago por corrida é de $6,73.

Explicação:

Nessa consulta eu comecei utilizando o WITH para fazer uma 'SubConsulta' chamada de 'dataByPassengers', uma SubConsulta basicamente serve para você fazer uma espécie de tratamento em alguns dados antes de utilizá-los na consulta principal. Nessa SubConsulta eu selecionei o 'passengerCount' e depois disso selecionei o 'tollsAmount' e fiz um tratamento usando o CAST() para passar ele para FLOAT64, utilizei o ROUND() para fixar um número de casas decimais e renomeei o campo para 'tools'.
Na linha de baixo fiz um COUNT(*) para captar o número de corridas e renomeei para 'totalRaces'.
Após isso fiz um filtro utilizando o WHERE onde os 'passengerCount' deveriam ser menores ou igual a '3' e o 'tollsAmount' deveriam ser diferentes de '0'.
Para finalizar nossa SubConsulta agrupei tudo por 'passengerCount'.
Começando nossa Consulta Principal, comecei ela somando todos os nosso campos 'tolls' anteriormente criado na SubConsulta 'dataByPassengers' e renomeando novamente para 'tolls'.
Também fiz a soma do campo 'totalRaces' que também foi criado na SubConsulta 'dataByPassengers', renomeie ele para 'totalRaces' novamente.
Por fim fiz a média dos pedágios por corrida, que era o solicitado no enunciado, para isso juntei a linha de código do 'tolls' sendo dividida pela linha de código do 'totalRaces' e renomeando para 'mediumTollsByRaces'.
Para finalizar fiz um FROM 'dataByPassengers' que era nossa SubConsulta.
Lembrando que todos os ROUND() usado na consulta foram para efeito de arredondamento das casas decimais.

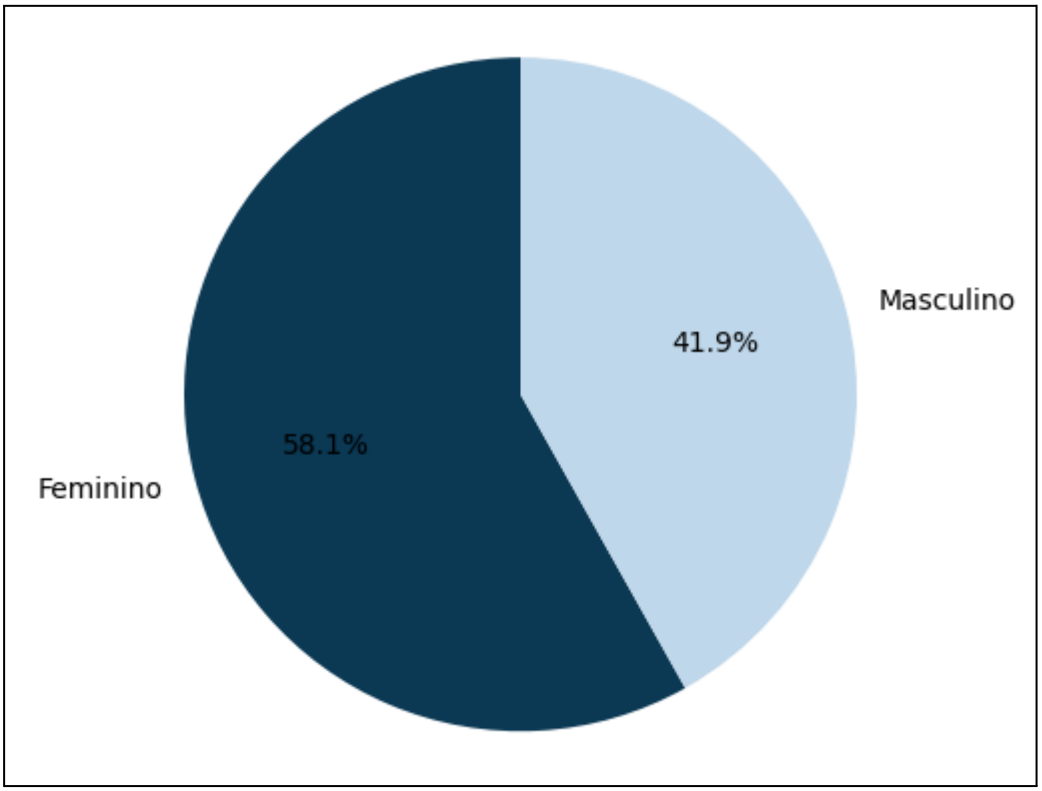
Projeto em Análise de Dados

- Análise de Casamentos Homoafetivos por Ano(Gráfico de Barras)



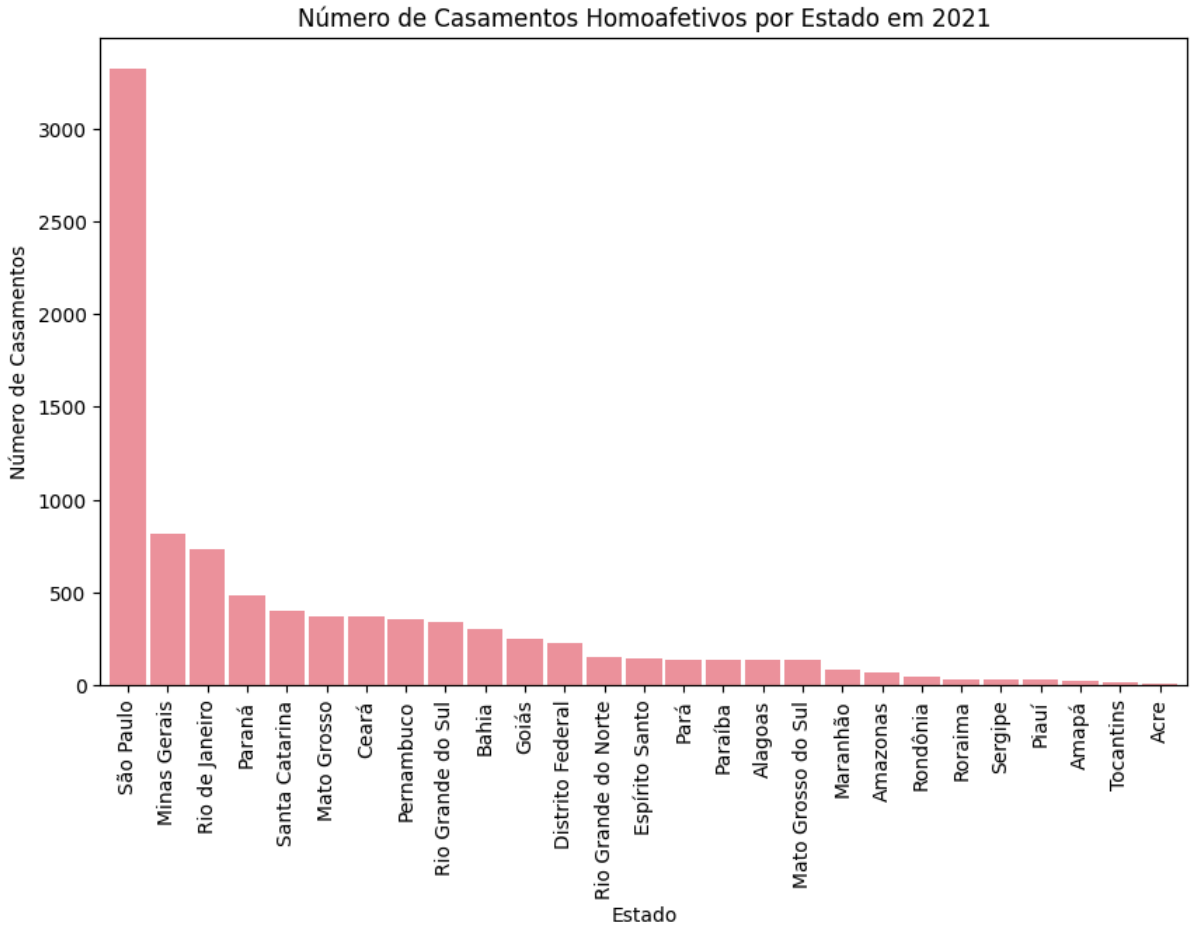
- Podemos notar que o volume de Casamentos Homoafetivos deram um grande salto a partir de 2018.
- Entre 2015 e 2017 tínhamos uma média de 5.618 casamentos por ano.
- Entre 2018 e 2019 nossa média de casamentos subiu para 9.288 por ano, um aumento de 65,32% comparado aos anos anteriores.
- Em 2020 notamos uma baixa nos casamentos que vinham em um crescimento, temos uma matéria do G1 que fala um pouco sobre as possíveis causas.
- Em 2021 já notamos um crescimento novamente, se equiparando com os anos de 2018 e 2019, devido às medidas mais frouxas do isolamento social.

- Análise de Casamentos Homoafetivos por Gênero(Gráfico de Pizza)



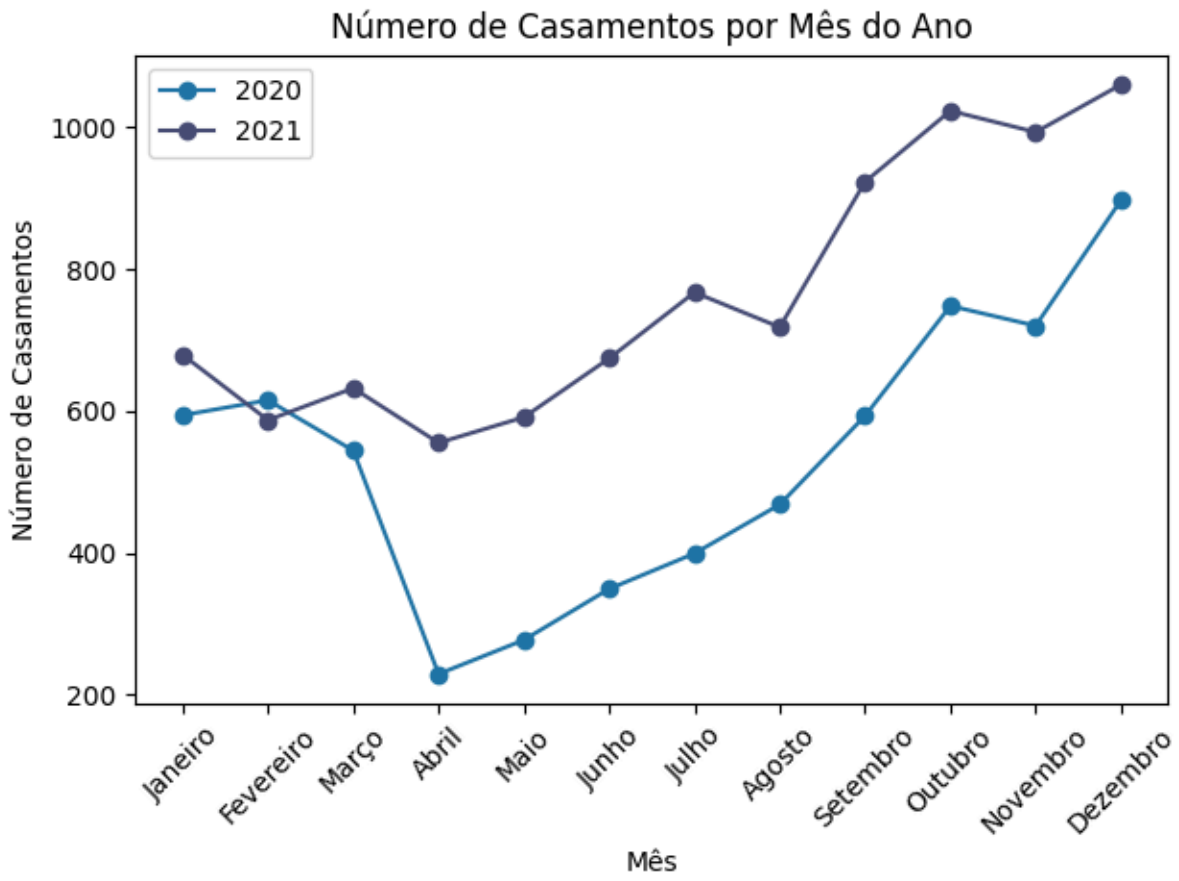
- Nessa visão podemos notar que ao longo dos anos, ocorreu uma procura maior de parceiras do sexo Feminino para se casar em cartorio.
- Temos uma matéria na Agência Brasil onde confirmam esses dados sobre os casamentos homoafetivos terem um número maior entre mulheres.
- Ainda nessa matéria também vemos que os dados apontam que o maior número de casamentos entre mulheres foi na Região Sudeste do Brasil.
- Já a menor foi a Região Norte do nosso País.

- Casamentos por Estado com base no Último Ano



- Podemos observar nesse gráfico que o maior estado a realizar casamentos homoafetivos é em São Paulo.
- Nenhum outro estado do Brasil se compara com a procura de casamentos que existe na Capital Paulista.
- Rio de Janeiro e Minas Gerais aparecem em segundo e terceiro lugar.
- Já na lanterna do ranking aparecem dois estados da região Norte do país: Tocantins e Acre.

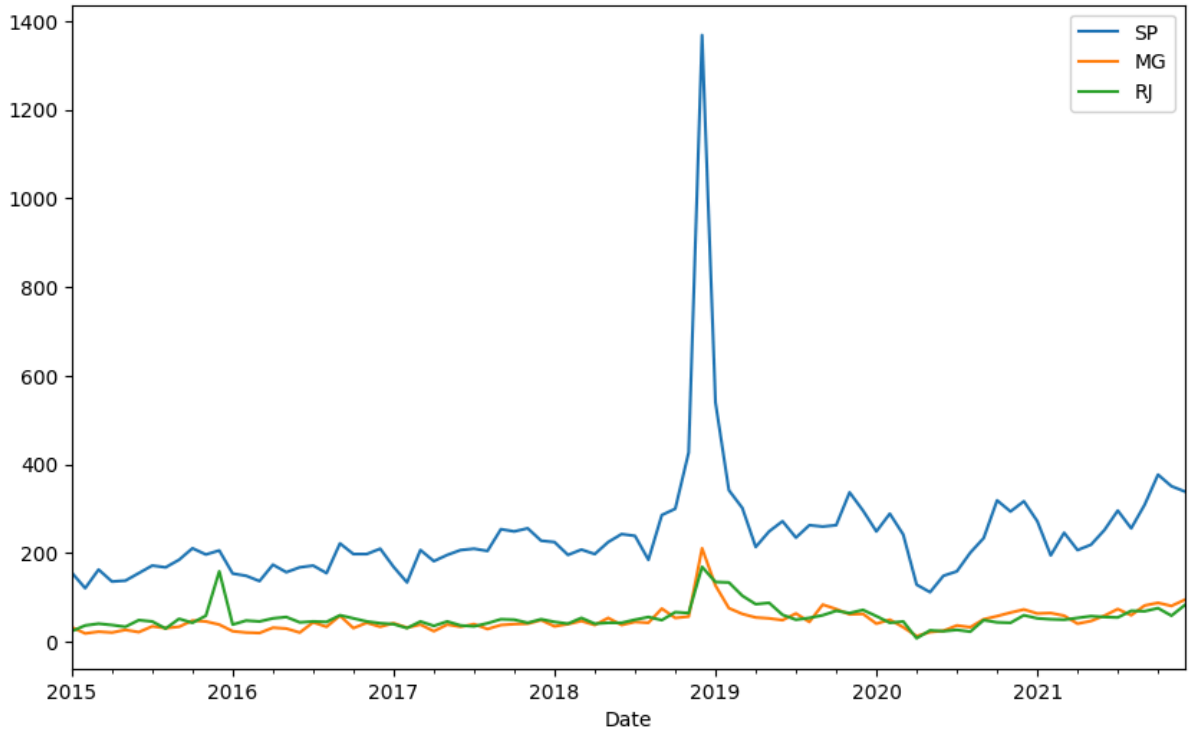
- Comparação de Casamentos por Mês dos ùltimos Dois Anos



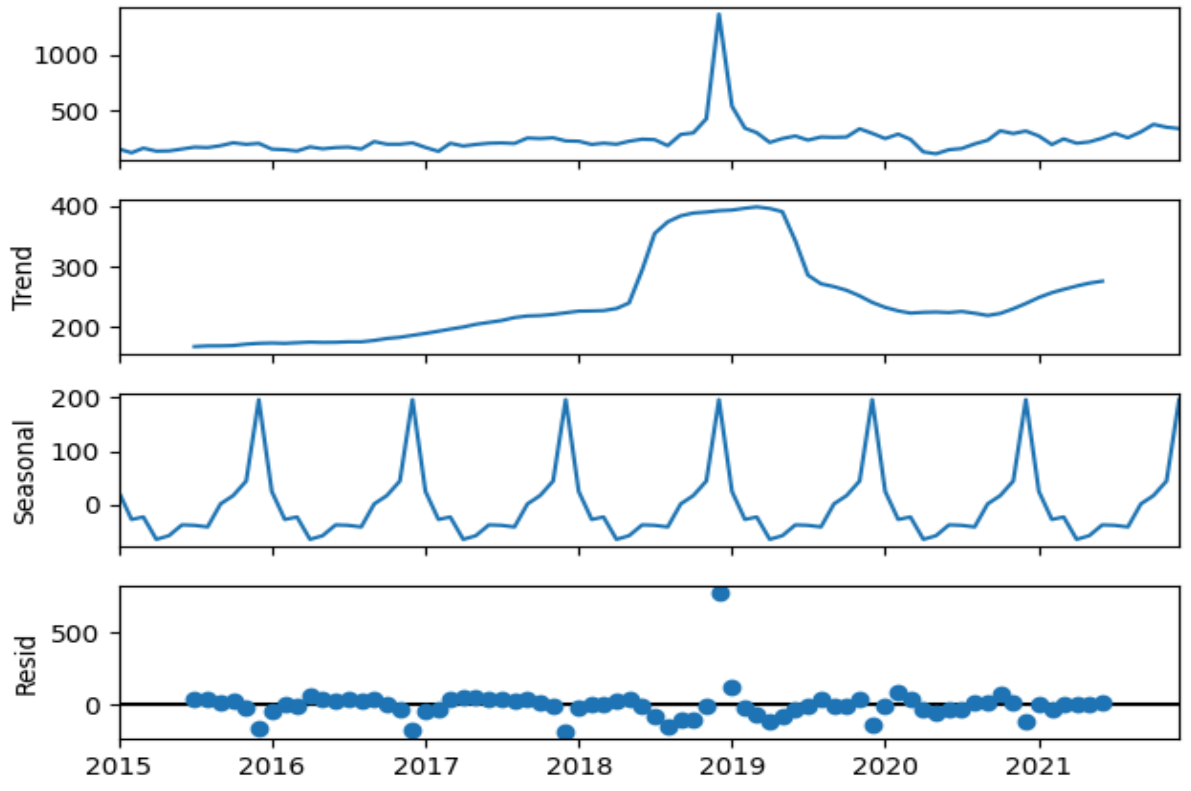
- Como podemos ver, começamos o ano de 2020 com um número significativo de casamentos, porém houve um declínio quando começou o isolamento social.
- No mês de outubro de 2020 já observamos o salto nos números, onde já ultrapassou o começo do mesmo ano.
- Já em 2021 vemos os dados onde aumentam e diminuem até o mês de maio.
- Após maio subiram os números até o mês 12 onde está no topo do gráfico.

Séries Temporais

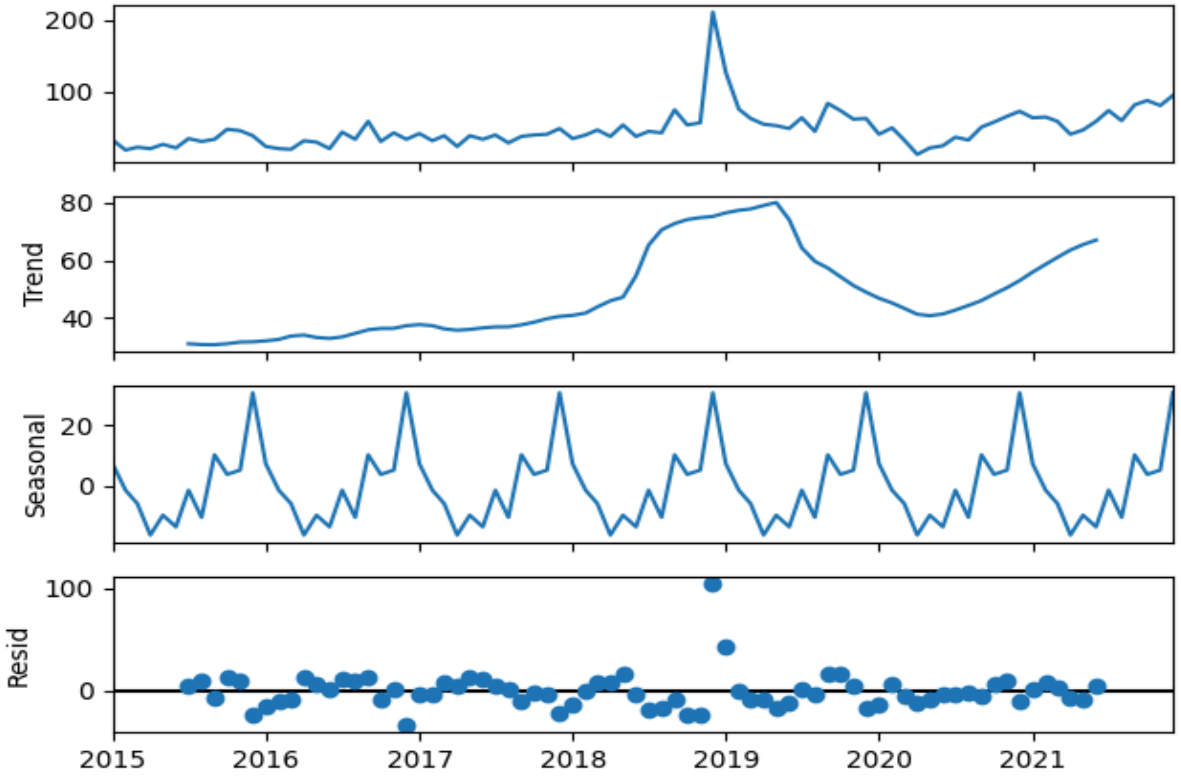
- Gráfico de Linhas baseado em Casamentos por mês dos 3 maiores Estados.



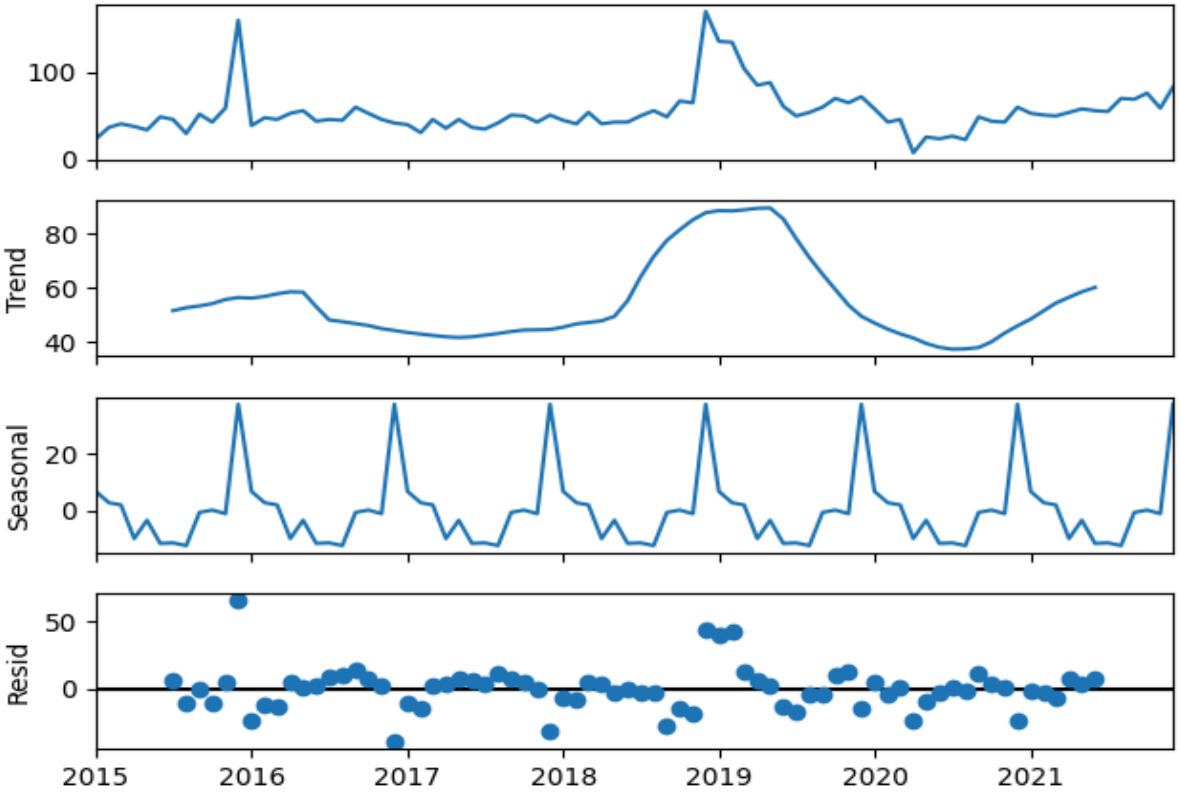
- Decomposição da Série Temporal em Gráficos de Tendência, Sazonalidade e Ruído do estado de São Paulo.



- Decomposição da Série Temporal em Gráficos de Tendência, Sazonalidade e Ruído do estado de Minas Gerais.

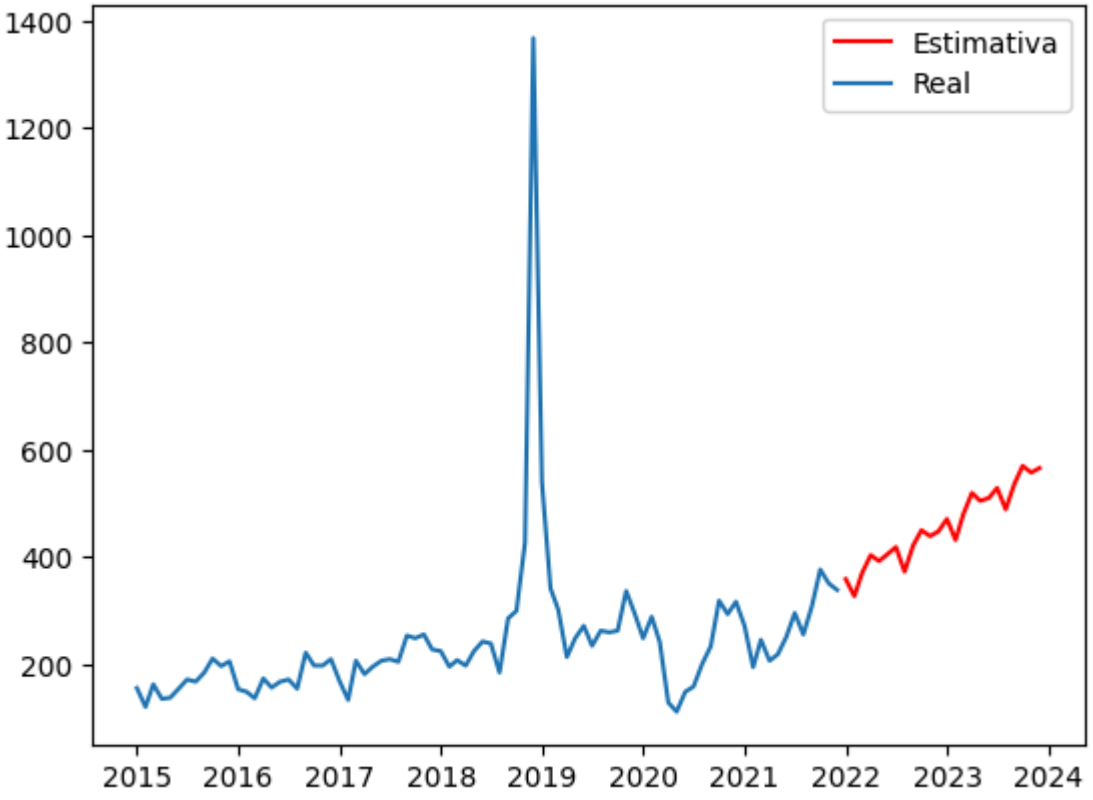


- Decomposição da Série Temporal em Gráficos de Tendência, Sazonalidade e Ruído do estado do Rio de Janeiro.

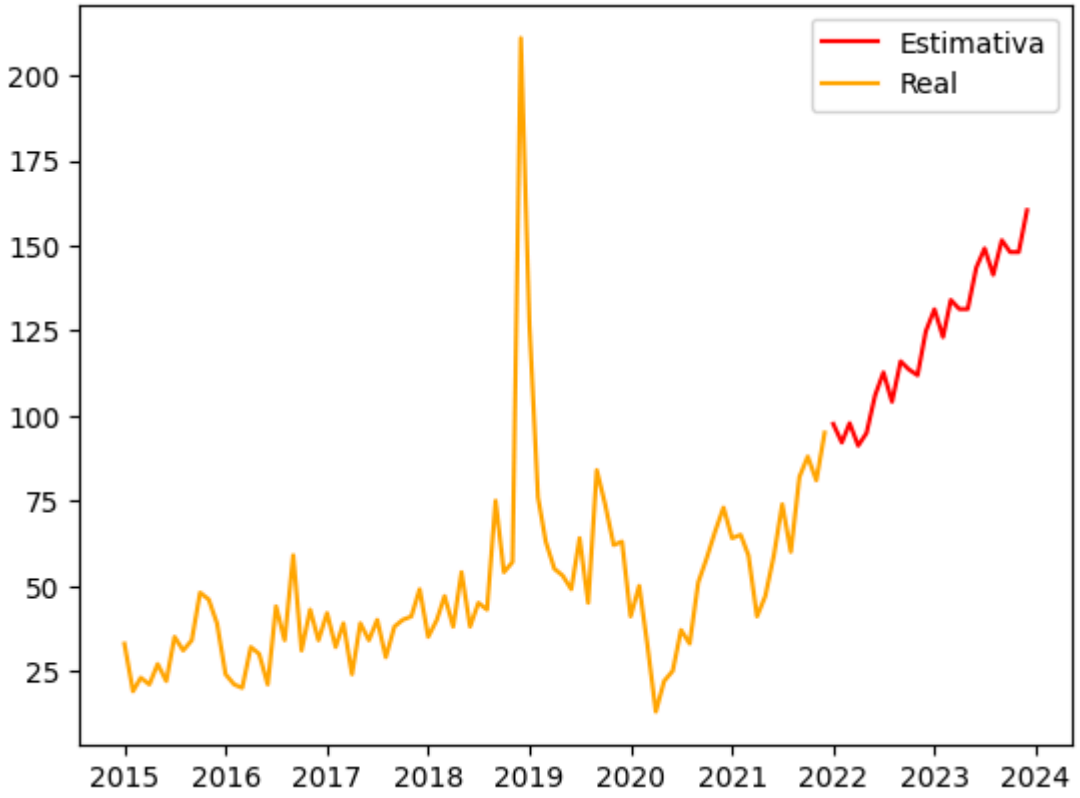


Previsões com ARIMA

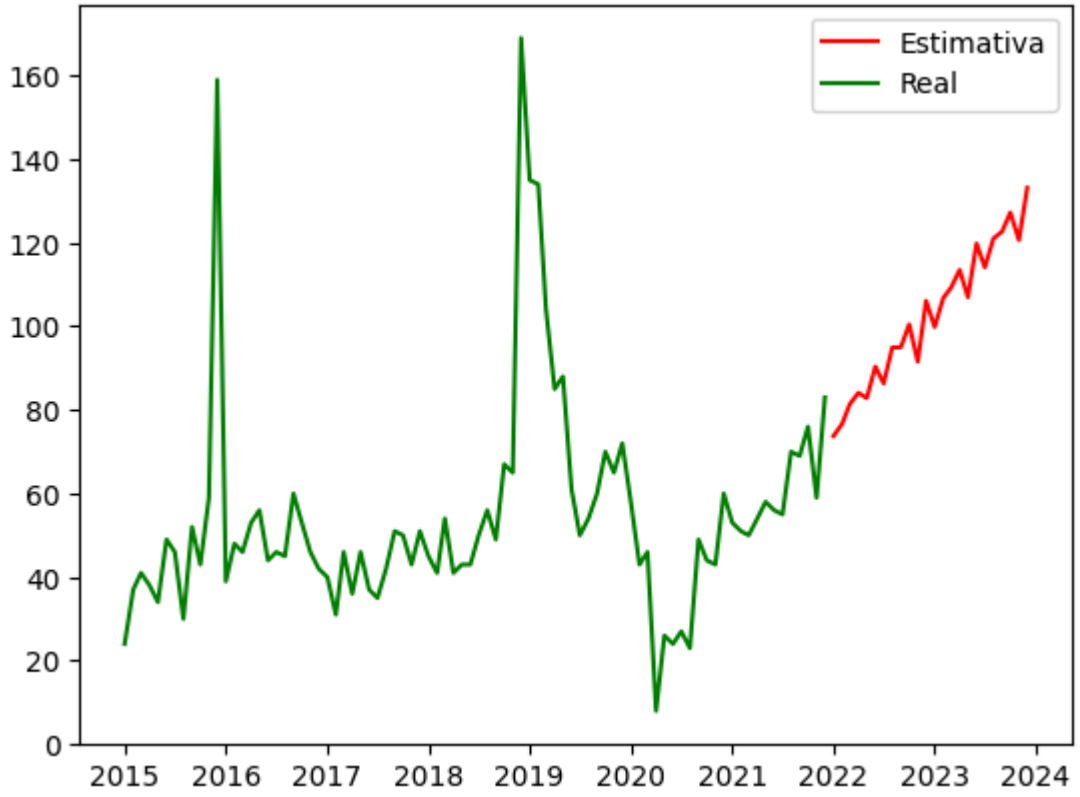
- Estimativa da Quantidade de Casamentos para os próximos 2 Anos no estado de São Paulo.



- Estimativa da Quantidade de Casamentos para os próximos 2 Anos no estado de Minas Gerais.



- Estimativa da Quantidade de Casamentos para os próximos 2 Anos no estado do Rio de Janeiro.



Conclusões do Projeto

Gostei bastante de ter a oportunidade de participar desse processo seletivo e também poder fazer esse Case, me diverti bastante fazendo ele, com ele tive vários aprendizados novos dos quais eu não tinha tido contato ainda, e com certeza isso me deu um gás para continuar nos estudos, independente do resultado desse projeto em si.
Séries temporais são uma destas coisas que ainda não tinha tido a oportunidade de ver e graças ao conteúdo do canal Let's Data foi possível eu desenvolver essa projeção dos próximos dois anos de Casamentos Homoafetivos.
Uma das coisas que me deixou feliz desenvolvendo esse projeto foi ver que consegui adquirir alguns conhecimentos rápido, tendo em vista que na última semana foi final de semestre e tive que focar o Case para esse último fim de semana, então entrego esse Case sentindo que ganhei muitos aprendizados novos e com o sentimento que estou evoluindo nessa área de DS, desde já agradeço pela oportunidade de fazer esse projeto, com certeza já ganhei muitas coisas com ele.