trabalho-gustavo-enxuto

April 2, 2019

1 Trabalho de análise multivariada

Nome: Gustavo Salvioli da Silva

```
In [1]: %matplotlib inline
        import pandas as pd
        import numpy as np
        import scipy.stats as stats
        import matplotlib.pyplot as plt
        import seaborn as sns
        sns.set(style='whitegrid')
        pd.set_option('mode.chained_assignment', None)
        # pacote altair para fazer outros tipos de gráficos
        # usaremos para o diagrama de dispersão com a cor
        import altair as alt

        # ACP
        from sklearn.decomposition import PCA
        from sklearn.preprocessing import scale
        # AA
        from scipy.cluster.hierarchy import dendrogram, linkage
        from scipy.spatial.distance import pdist
        from scipy.cluster.hierarchy import cut_tree
        from sklearn.cluster import KMeans
```

Neste trabalho serão analisados dados do Ministério da Agricultura, Pecuária e Abastecimento, em 2017, referentes ao Programa de Subvenção ao Prêmio proposto pelo governo Federal, afim de auxiliar no desenvolvimento do seguro Rural no Brasil. Observando os valores totais dos prêmios pagos em 2017, divididos em cultura, para cada mesorregião do Brasil. As culturas serão soja, milho1(safra), milho2(safrinha), pecuária, cana, trigo, floresta, feijão, uva, café, tomate e arroz. O prêmio é a importância paga por alguém a uma seguradora em troca da transferência do risco á que ele está exposto.

1.1 Ler o conjunto de dados csv (*dataframe*)

```
In [ ]: from google.colab import files
        uploaded = files.upload()
```

```
<IPython.core.display.HTML object>


Saving segurorural-meso.csv to segurorural-meso.csv
```

```
In [2]: dados = pd.read_csv('/home/patricia/drive/unifal/analise-multivariada/2018-2/trabalho/se
```

O *dataframe* dados2 inclui as mesorregiões para usarmos posteriormente nas análises. Por enquanto não o utilizaremos.

```
In [5]: dados.describe()
```

Out[5]:

	soja	milho1	milho2	pecuaria	cana
count	1.250000e+02	1.250000e+02	1.250000e+02	125.0000	125.00000
mean	3.503304e+06	1.313973e+05	1.269308e+06	11130.1680	36618.56000
std	6.876624e+06	3.019043e+05	4.006011e+06	34387.8348	112733.00222
min	0.000000e+00	0.000000e+00	0.000000e+00	0.0000	0.00000
25%	6.445200e+04	0.000000e+00	0.000000e+00	0.0000	0.00000
50%	4.792660e+05	2.155800e+04	2.418700e+04	0.0000	0.00000
75%	3.073415e+06	1.047780e+05	3.834520e+05	6884.0000	11228.00000
max	3.507858e+07	2.463508e+06	2.798412e+07	276834.0000	819811.00000

	trigo	floresta	feijao	uva	cafe
count	1.250000e+02	125.000000	1.250000e+02	1.250000e+02	1.250000e+02
mean	5.597157e+05	23947.024000	7.660636e+04	4.635794e+05	1.036416e+05
std	1.963048e+06	66354.869181	3.334532e+05	3.284582e+06	4.777835e+05
min	0.000000e+00	0.000000	0.000000e+00	0.000000e+00	0.000000e+00
25%	0.000000e+00	0.000000	0.000000e+00	0.000000e+00	0.000000e+00
50%	0.000000e+00	0.000000	0.000000e+00	0.000000e+00	0.000000e+00
75%	7.875500e+04	15935.000000	9.362000e+03	6.935000e+03	5.847000e+03
max	1.522280e+07	528221.000000	2.932674e+06	3.550678e+07	4.664392e+06

	tomate	arroz
count	1.250000e+02	1.250000e+02
mean	1.329638e+05	2.060330e+05
std	5.740098e+05	9.161344e+05
min	0.000000e+00	0.000000e+00
25%	0.000000e+00	0.000000e+00
50%	0.000000e+00	0.000000e+00
75%	3.063000e+03	0.000000e+00
max	5.363316e+06	6.131478e+06

Interpretação: São 125 observações, ou seja, as mesorregiões do Brasil, são analisadas em cada uma das mesorregião o valor do prêmio total pago em 2017, por alguns produtos agrícolas.A soja tem a maior média do prêmio pago. Todas as culturas tem alguma mesorregião que não contratou o seguro para alguma modalidade. A soja é a única cultura que tem o primeiro quantil e tem a maior mediana, significando que 75% das mesorregiões tiveram pagamentos de prêmios para essa modalidade. Nas cultura de soja, milho1 e milho2 apresentam uma mediana maior que a média,

ou seja os valores dos prêmios pagos por mesorregião são mais parecidos diferentemente das outras culturas que a média é maior, tendo assim prêmios pagos por mesorregião mais diferentes um dos outros. Em relação ao segundo quartil somente as culturas milho1 e milho2 tiveram o quartil diferente de zero, demonstrando que, 50 % das mesorregiões tiveram pagamentos dos prêmios para uma dessas culturas . Já os produtos agrícolas pecuária, cana, trigo, floresta, feijão, uva e café tiveram apenas o terceiro quartil, dessa forma, 25 % das mesorregiões do Brasil tiveram pagamento do prêmio na contratação do seguro rural para alguma dessas culturas. E o produto agrícola arroz, menos de 25 % das mesorregiões tiveram pagamento do prêmio relacionado a essa cultura.

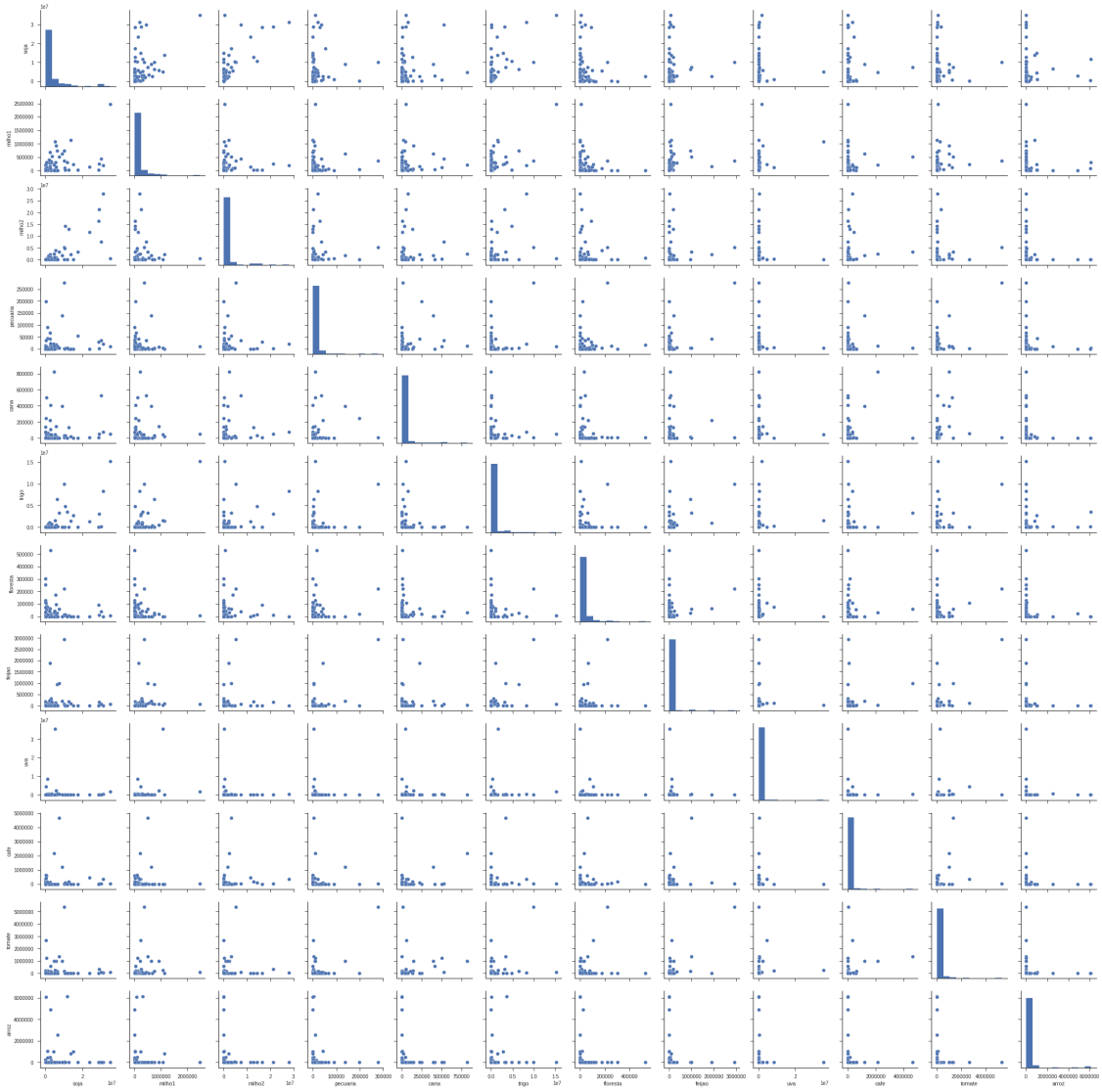
```
In [ ]: dados.corr()
```

```
Out[ ]:                soja    milho1    milho2  pecuaria      cana     trigo  \
        soja      1.000000  0.520024  0.746524  0.151412  0.210765  0.613071
        milho1    0.520024  1.000000  0.051094  0.079842  0.146748  0.655072
        milho2    0.746524  0.051094  1.000000  0.105939  0.135468  0.385292
        pecuaria  0.151412  0.079842  0.105939  1.000000  0.220828  0.292932
        cana      0.210765  0.146748  0.135468  0.220828  1.000000 -0.003818
        trigo     0.613071  0.655072  0.385292  0.292932 -0.003818  1.000000
        floresta  0.023258 -0.034737  0.039272  0.203699 -0.014015  0.082660
        feijao    0.119925  0.188979  0.111033  0.585571  0.068518  0.465096
        uva       0.031952  0.322024 -0.030079 -0.023165  0.009450  0.076756
        cafe      0.098316  0.130206  0.116453  0.045539  0.284987  0.114114
        tomate    0.103951  0.184463  0.093213  0.611246  0.240270  0.385366
        arroz     0.064511  0.008548 -0.064035 -0.047791 -0.069846  0.059302

                  floresta    feijao       uva      cafe    tomate     arroz
        soja      0.023258  0.119925  0.031952  0.098316  0.103951  0.064511
        milho1   -0.034737  0.188979  0.322024  0.130206  0.184463  0.008548
        milho2    0.039272  0.111033 -0.030079  0.116453  0.093213 -0.064035
        pecuaria  0.203699  0.585571 -0.023165  0.045539  0.611246 -0.047791
        cana     -0.014015  0.068518  0.009450  0.284987  0.240270 -0.069846
        trigo     0.082660  0.465096  0.076756  0.114114  0.385366  0.059302
        floresta  1.000000  0.245784 -0.003947  0.053313  0.264451 -0.063217
        feijao    0.245784  1.000000 -0.007341  0.221923  0.705900 -0.049303
        uva      -0.003947 -0.007341  1.000000 -0.009935  0.083516 -0.021498
        cafe      0.053313  0.221923 -0.009935  1.000000  0.274887 -0.047722
        tomate    0.264451  0.705900  0.083516  0.274887  1.000000 -0.049879
        arroz    -0.063217 -0.049303 -0.021498 -0.047722 -0.049879  1.000000
```

Interpretação: Mostra a correlação dos prêmios pagos em 2017 entre as culturas pelas mesorregiões, podendo notar a maior correlação positiva entre a soja e o milho2(0.75), seguida do tomate e feijão(0.70).

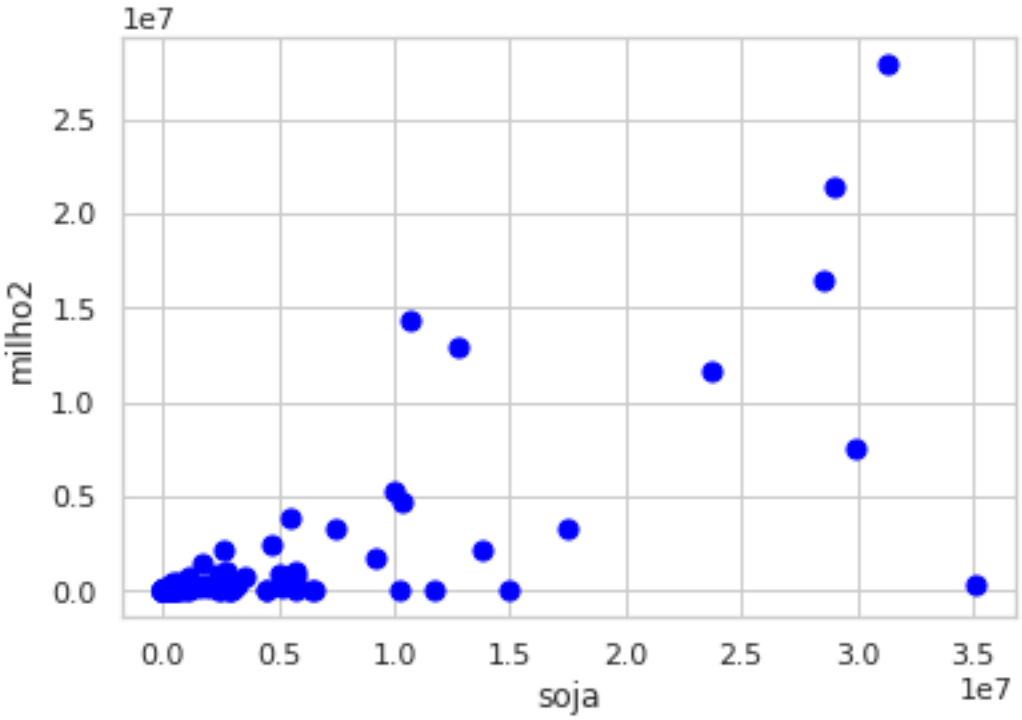
```
In [ ]: sns.set(style='ticks')
        sns.pairplot(dados);
```



Interpretação: Pode- se observar no gráfico as relações entre uma cultura e outra, pelas mesorregiões e chegar em conclusões que; o milho2 e a soja é a que apresenta uma maior correlação alta e positiva, podemos supor que , essas mesorregões contratam seguros para o milho2 e para a soja, desta maneira, a contratação para uma cultura poderá haver também contratação para outra. Outra correlação alta e positiva está nas culturas de tomate e feijão, a qual podemos imaginar que algumas messoregiões á uma demanda pelas contratações de seguros das duas culturas, e que talvez os preços dos prêmios pagos são parecidos, devido ao grau de risco que as seguradoras enfrentam pelas indenizações pagas das produções. Outras correlações alta e positiva no prêmio total pago entre as mesorregiões está no trigo e milho1, trigo e soja, podendo supor as mesmas idéias das relações acima. Uma outra relação que apresenta a correlação mais fraca está nas culturas entre cana e trigo, sendo negativa, e a correlação mais negativa está entre arroz e cana, podendo supor, consequentemente aonde á uma cultura em uma mesorregião não á em outra, devido a condições distintas a qual as culturas necessitam para crescer. Analisando as correlações das culturas relacionadas as somas dos prêmios pagos por mesorregião em 2017, podemos sus-

peitar de que, as mesorregiões parecidas em condições da natureza e econômicas, apresentam uma contratações para os seguros parecidas para as mesmas culturas como o caso, do milho2 e a soja. E devido a essas demandas podemos, ver quais as mesorregiões pode estar produzindo quais culturas, podendo suspeitar dessa forma, os prêmios pagos por mesorregião, por culturas pode-se haver uma relação direta entre produções agrícolas por mesorregião e contratações de seguros para esses produtos agrícolas.

```
In [7]: dados.plot.scatter('soja', 'milho2', s=50, c='blue');    # há opções (usar SHIFT TAB)
```




```
In [36]: ## ACP usando a matriz de correlações
         # apenas mudando o nome do conjunto de dados para X e desconsiderando o código
         X = dados.iloc[:, :]
         # efetua a ACP
         pca = PCA()
         resultado_pca = pca.fit_transform(scale(X))    # scale(X) padroniza os dados, como se us
         # salva os escores dos dois primeiros CPs em um dataframe
         resultado = pd.DataFrame({'cp1':resultado_pca[:, 0], 'cp2':resultado_pca[:, 1]}, index=
         # coeficientes dos CPs (autovetores)
         # cada linha é um CP
         # $Y_1$: as variáveis que mais influenciam no CP1 são 'trigo', 'tomate' e 'feijão' de u
         # $Y_2$: as variáveis ao qual mais influenciam no CP2 são 'soja', de forma negativa, 't
         # $Y_3$: as variáveis que mais enfluenciam no CP3 são 'uva', de forma positiva e 'milho
         # variância explicada acumulada
         # cria um dataframe com os resultados da ACP
```

```
        pca_df = pd.DataFrame(
            resultado_pca,
            index=X.index,
            columns=['CP' + str(i + 1) for i in range(resultado_pca.shape[1])]
        )
        # loadings de cada componente
        loadings = pd.DataFrame(
            pca.components_,
            index=['CP'+str(i+1) for i in range(len(pca.components_))],
            columns=X.columns
        ).T
```

```
/home/patricia/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:6: DataConversionWarn
```

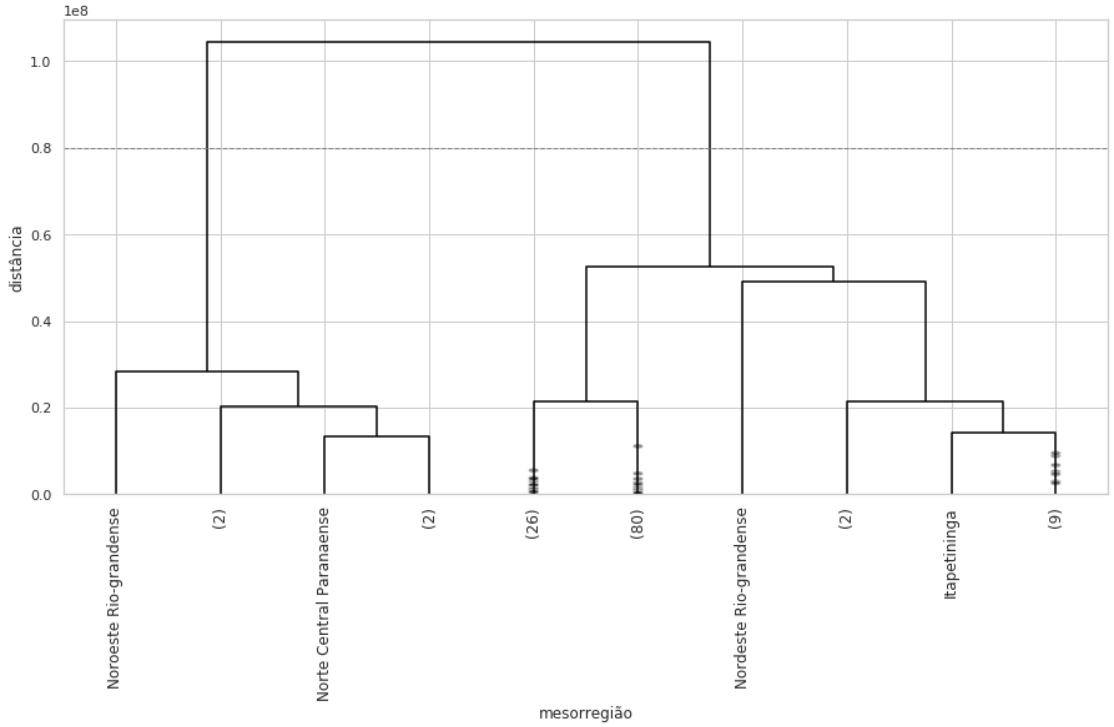
1.2 Análise de agrupamento

Ward

```
In [37]: Z = linkage(X, method='ward')
```

```
In [38]: # definir a distância de corte baseando no dendrograma
         max_d = 80000000
         grupos = cut_tree(Z, height=max_d)
```

```
In [39]: # dendrograma com mais opções
         # mostra o ponto de corte na distância max_d
         fig, ax = plt.subplots(figsize=(15, 7))
         ax = dendrogram(
             Z,
             truncate_mode='lastp',  # mostrar apenas os p últimos grupos formados
             p=10,  # quantos passos mostrar
             show_leaf_counts=True,  # mostrar quantas observações há em cada grupo entre parênt
             leaf_rotation=90., # rotação
             leaf_font_size=12., # tamanho da fonte
             labels=dados.index, # rótulos do eixo x
             show_contracted=True,  # to get a distribution impression in truncated branches,
             above_threshold_color='black',
             color_threshold=0.1, # para que todas as linhas sejam da mesma cor
             # color_threshold=max_d, # para que os grupos fiquem com cores diferentes
         )
         plt.axhline(y=max_d, c='grey', lw=1, linestyle='dashed')
         plt.xlabel('mesorregião')
         plt.ylabel('distância');
```



O método Ward, separou os outliers, como nos outros métodos, mas tendeu a balancear um pouco melhor. Dessa forma, o número de grupos escolhidos foram 2 grupos, devido a interpretabilidade entre os grupos.

Método escolhido:

```
In [40]: Z = linkage(X, method='ward')
         max_d = 80000000
         grupos = cut_tree(Z, height=max_d)
         # incluir no resultado dos escores dos dois primeiros CPs a informação sobre os grupos
         resultado['grupo'] = grupos
         # contagem de observações em cada grupo
         resultado.grupo.value_counts()

Out[40]: 0    119
         1      6
         Name: grupo, dtype: int64

In [41]: # incluir no dataframe de dados as informações sobre a qual grupo cada observação perte
         dados['grupo'] = grupos

In [42]: # média dos grupos - todas as variáveis
         # inclusive as não utilizadas para agrupar
         dados.groupby('grupo').mean()

Out[42]:                soja         milho1        milho2      pecuaria           cana  \
         grupo
```

```
0      2.188174e+06  108655.840336  6.159729e+05  10873.957983   32490.848739
1      2.958672e+07  582436.333333  1.422712e+07  16211.666667  118484.833333

               trigo      floresta        feijao            uva           cafe  \
grupo
0      3.510477e+05  23908.168067  77924.016807  470164.302521  101550.478992
1      4.698299e+06  24717.666667  50472.833333  332977.833333  145115.666667

               tomate          arroz    grupos
grupo
0      135075.613445  216258.420168  2.806723
1       91079.833333    3229.833333  3.000000
```

```
In [43]: # mediana das variáveis para cada grupo
         dados.groupby('grupo').median()
```

```
Out[43]:              soja    milho1      milho2  pecuaria     cana      trigo  \
         grupo
         0        410688.0   20715.0     22316.0       0.0      0.0        0.0
         1      29457363.5  220389.0  14042101.0   15557.0  51360.5  2230023.0

                floresta   feijao     uva     cafe   tomate  arroz  grupos
         grupo
         0           0.0      0.0     0.0      0.0      0.0    0.0       1
         1        9783.5  36045.0  7848.5  27766.0  55636.0    0.0       3
```

Utilizando o método Ward, a divisão das mesorregiões do Brasil ficou um 'pouco' mais balanceada separando os outliers, porém ainda haum grupo com grande partes das mesorregiões.

- O grupo 0, tem maiores médias de prêmios para as culturas de feijão, uva, tomate e arroz, porém suas medianas são igual a zero, dessa forma, pode-se dizer que metade das mesorregiões do Brasil em 2017 não tiveram pagamentos dos prêmios para algumas dessas culturas, portanto algumas dessas mesorregiões em que houveram os pagamentos desses prêmios foram valores muito altos influenciando assim na média.
- Em relação ao grupo 1, obteve maiores médias de prêmios para as culturas agrícolas de soja, milho1, milho2, pecuária, cana, trigo, floresta e café. E analisando as medianas, pode-se dizer, que nesse grupo pelo menos 50% das mesorregiões pagaram prêmios para alguma dessas culturas, sendo a única exceção o arroz.

1.3 As observações de cada grupo

```
In [44]: grupo0 = dados.query('grupo == 0').index
         # list(grupo0)
         grupo1 = dados.query('grupo == 1').index
         # list(grupo1)
```

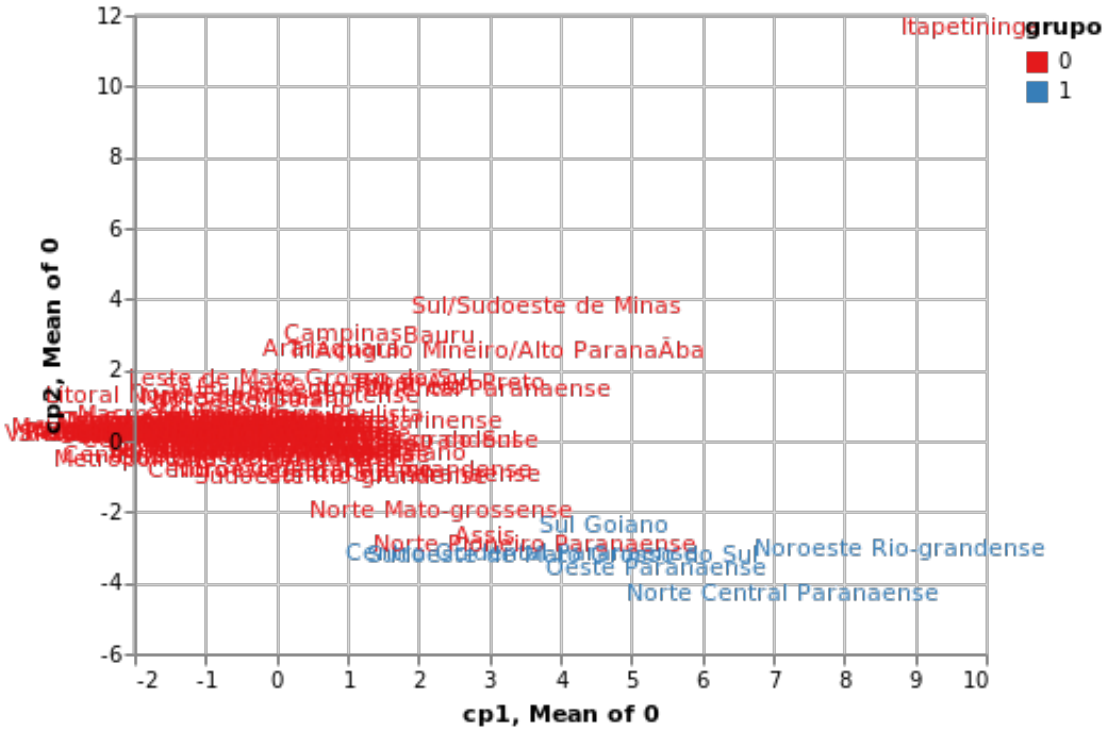
```
In [45]: resultado['nome_meso'] = resultado.index
```

In [46]:
```
# gráficos
g = alt.Chart(resultado).mark_text().encode(
    alt.X('cp1', scale=alt.Scale(domain=[resultado.cp1.min(), resultado.cp1.max()])),
    alt.Y('cp2', scale=alt.Scale(domain=[resultado.cp2.min(), resultado.cp2.max()])),
    text='nome_meso',
    color=alt.Color('grupo:O', scale=alt.Scale(scheme='set1'))
)
# eixo x = 0
bar_x = alt.Chart(resultado).mark_rule(color='gray').encode(
    x='mean(0):Q')

# eixo y = 0
bar_y = alt.Chart(resultado).mark_rule(color='gray').encode(
    y='mean(0):Q')


g + bar_x + bar_y
```

Out[46]:



1.4 K-médias

1.4.1 Gráfico k x SQDG

O SQDG sugere que a escolha do número de grupos (k) deveria ser 4, porém pela interpretabilidade do número de grupos foram escolhidos 2 grupos.

```
In [47]: # número de grupos sugerido pelo dendrograma
         k = 2
         kmeans = KMeans(n_clusters=k, random_state=10).fit(X)
         # incluir no resultado dos escores dos dois primeiros CPs a informação sobre os grupos
         # com o método das k-médias
         resultado['grupo'] = kmeans.labels_
         # contagens
         resultado.grupo.value_counts()

Out[47]: 0    116
         1      9
         Name: grupo, dtype: int64

In [48]: # incluir no dataframe de dados as informações sobre a qual grupo cada município perten
         dados['grupo'] = kmeans.labels_

In [49]: # média dos grupos - todas as variáveis
         # inclusive as não utilizadas para agrupar
         dados.groupby('grupo').mean()

Out[49]:                soja         milho1        milho2      pecuaria          cana  \
         grupo
         0      1.892727e+06  110947.017241  3.686735e+05  10674.181034  31928.301724
         1      2.426186e+07  394978.777778  1.287748e+07  17007.333333  97070.777778

                       trigo      floresta        feijao            uva           cafe  \
         grupo
         0      3.184854e+05  24353.189655  78376.801724  482082.267241  101815.387931
         1      3.668907e+06  18712.000000  53787.333333  225097.333333  127179.555556

                       tomate          arroz    grupos
         grupo
         0      138568.948276  221851.310345  2.706897
         1       60719.888889    2153.222222  4.222222

In [50]: # mediana das variáveis para cada grupo
         dados.groupby('grupo').median()

Out[50]:              soja    milho1      milho2  pecuaria     cana      trigo  \
         grupo
         0        382526.5   19886.5     13511.0       0.0      0.0        0.0
         1      28594753.0  133370.0  12938019.0    9864.0  50847.0  1334068.0
```

```
        floresta  feijao     uva     cafe  tomate  arroz  grupos
grupo
0            0.0     0.0     0.0      0.0     0.0    0.0       1
1         8368.0  7067.0  5681.0  41641.0     0.0    0.0       3
```

Apresentando a maior diferença em relação ao grupo 1, ao qual, a única exceção era o arroz, ao qual ele tinha uma mediana igual zero, agora o tomate também apresenta uma mediana igual a zero, significando, que pelo menos 50% das mesorregiões não tem pagamento de prêmio, relacionado a alguma dessas culturas.

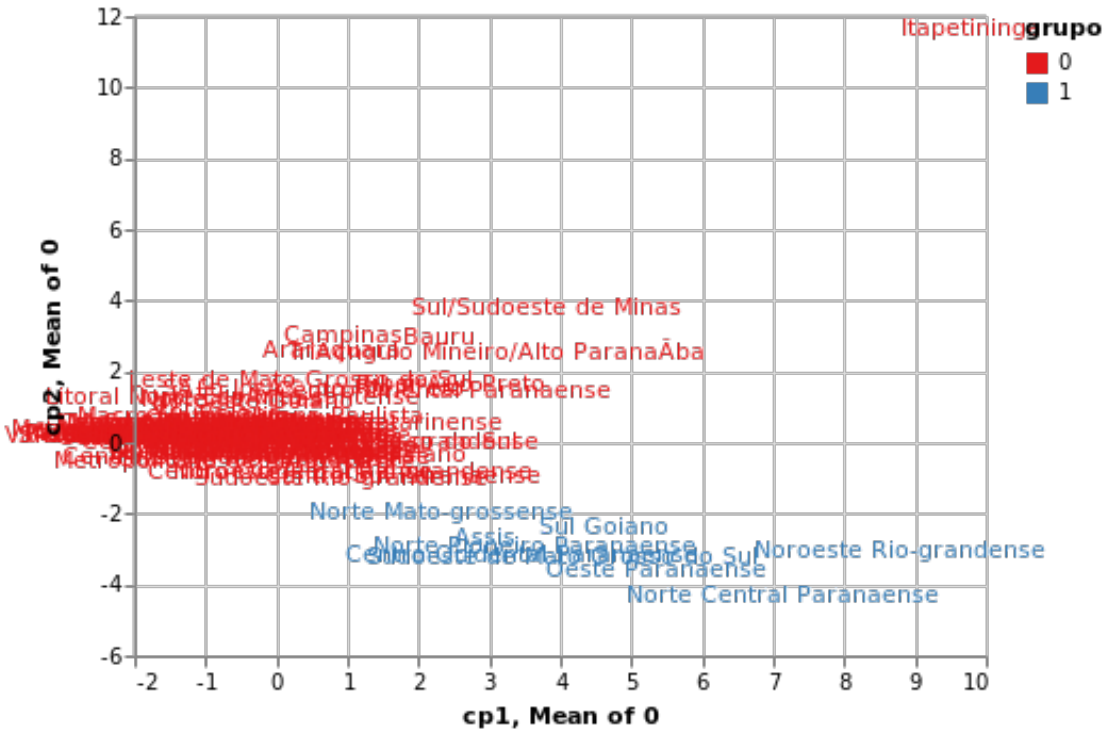
In [51]:
```
resultado['nome_meso'] = resultado.index
```

In [52]:
```
# gráficos
g = alt.Chart(resultado).mark_text().encode(
    alt.X('cp1', scale=alt.Scale(domain=[resultado.cp1.min(), resultado.cp1.max()])),
    alt.Y('cp2', scale=alt.Scale(domain=[resultado.cp2.min(), resultado.cp2.max()])),
    text='nome_meso',
    color=alt.Color('grupo:O', scale=alt.Scale(scheme='set1'))
)
# eixo x = 0
bar_x = alt.Chart(resultado).mark_rule(color='gray').encode(
    x='mean(0):Q')

# eixo y = 0
bar_y = alt.Chart(resultado).mark_rule(color='gray').encode(
    y='mean(0):Q')


g + bar_x + bar_y
```

Out[52]:



O método Ward e o método do K-médias apresentaram resultados parecidos com a suas maiores diferenças sendo no número de mesorregiões separadas entre os grupos, o grupo 1 no método Ward apresentou 6 mesorregiões e o método K-médias no grupo 1 apresentou 9 mesorregiões.

In [ ]: