

Adatbányászat a Gyakorlatban

5. Gyakorlat: Gyakorisági adatok kezelése

Kuknyó Dániel

Budapesti Gazdasági Egyetem

2024/25

1.félév

1 Bevezetés

1 Bevezetés

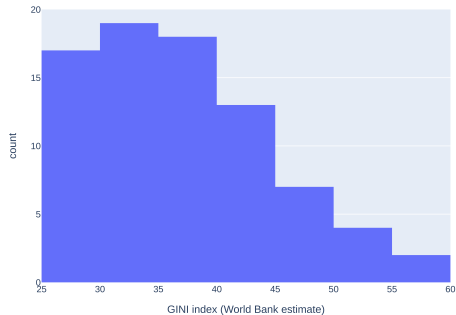
Hisztogramok létrehozása

Hisztogram

A hisztogram egy statisztikai grafikon, amely az adatok eloszlását mutatja be. Oszlopdiagram formájában ábrázolja, hogy az adatok milyen gyakorisággal fordulnak elő különböző intervallumokban.

Hisztogram létrehozása plotly segítségével:

```
1 px.histogram(data_frame=df, x=gini)
```



Hisztogramok felbontása

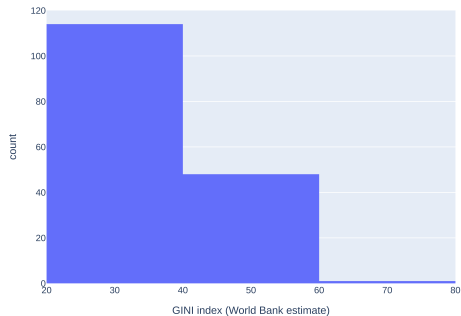
Osztályköz

Az osztályközök határozzák meg, hogy az adatok milyen tartományokba kerülnek, és ezek az intervallumok határozzák meg a hisztogram oszlopainak szélességét.

Az osztályközök száma az `nbins` paraméter segítségével állítható.

```
1 for n in [2, 45, 500]:  
2     px.histogram(data_frame=df, x=gini,  
                  nbins=n)
```

`nbins = 2`



Hisztogramok felbontása

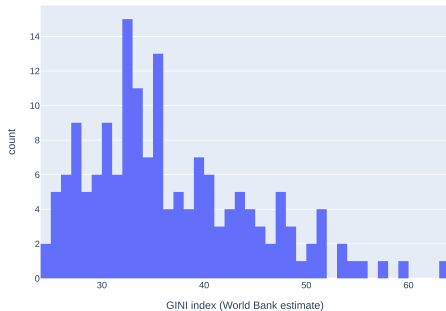
Osztályköz

Az osztályközök határozzák meg, hogy az adatok milyen tartományokba kerülnek, és ezek az intervallumok határozzák meg a hisztogram oszlopainak szélességét.

Az osztályközök száma az `nbins` paraméter segítségével állítható.

```
1 for n in [2, 45, 500]:  
2     px.histogram(data_frame=df, x=gini,  
                  nbins=n)
```

nbins = 45



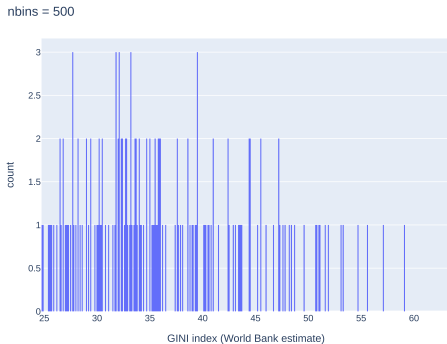
Hisztogramok felbontása

Osztályköz

Az osztályközök határozzák meg, hogy az adatok milyen tartományokba kerülnek, és ezek az intervallumok határozzák meg a hisztogram oszlopainak szélességét.

Az osztályközök száma az `nbins` paraméter segítségével állítható.

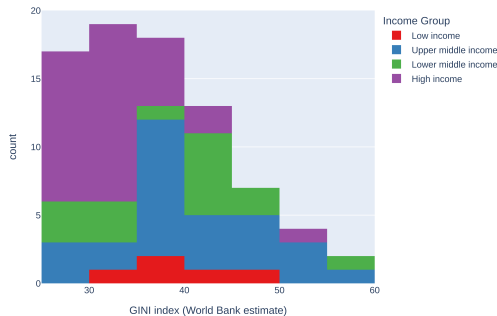
```
1 for n in [2, 45, 500]:  
2     px.histogram(data_frame=df, x=gini,  
                  nbins=n)
```



Hisztogram hasítása színekkel

Plotly express diagramokat lehetséges változón belüli csoportonként meghasítani. Ennek eléréséhez a `color` paramétert kell a megfelelő változóra állítani.

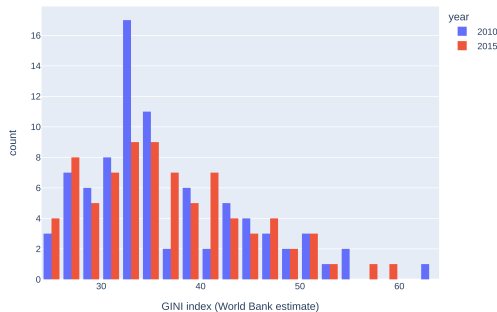
```
1 px.histogram(data_frame=df, x=gini,
               color='Income Group',
               color_discrete_sequence=px.colors.
               qualitative.Set1)
```



Csoportosított hisztogramok

Vannak olyan esetek, amikor egy változónak több csoportját egymás mellett szükséges megmutatni. Ekkor a hisztogramokat lehetséges csoportosítani adott értékek szerint, a `color` és a `barmode='group'` paraméterek állításával.

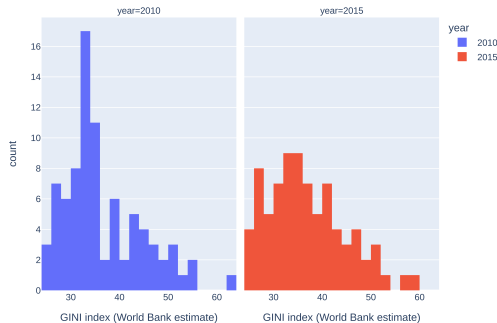
```
1 px.histogram(df, x=gini, color='year',  
    barmode='group')
```



Hasított hisztogramok

A diagramok hasítása adott változó értékei szerint lehetséges úgy is, hogy minden, a változóhoz tartozó értékre szűrt adathalmaz egy külön diagramon jelenik meg, a `facet_col` paraméter állításával.

```
1 px.histogram(df, x=gini, color='year',  
               facet_col='year')
```

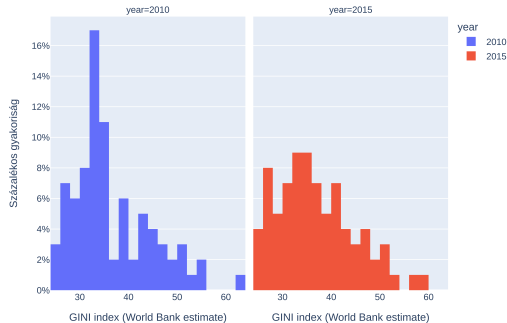


Hisztogramok normalizálása

Normalizált hisztogram

Olyan grafikon, ahol az egyes oszlopok az adott intervallumba eső adatok gyakoriságát jelzi olyan módon, hogy az oszlopok összege 1 legyen.

```
1 fig = px.histogram(df, x=gini, color='  
   year', facet_col='year')  
2 fig.layout.yaxis.ticksuffix = '%'  
3 fig.layout.yaxis.title = 'Százalékos  
   gyakoriság'  
4 fig.show()
```



Alkalmazás interaktív hisztogramokkal (freq_app_v1.py)

A callback függvények a felhasználói interakciók alapján frissítik a grafikonokat. A `display_histogram` függvény három bemeneti elemet figyel (years, indicator, bins), és ezek alapján frissíti a hisztogram ábrát.

Ha nincs kiválasztott év vagy indikátor, a függvény nem frissíti az ábrát (`PreventUpdate`). Az adatok szűrése után a Plotly Express segítségével készül el a hisztogram.

