Exploratory Data Analysis with polars Library

Alier Reng

Load Libraries

```
import polars as pl
  import polars.selectors as cs
  from plotnine import *
  from mizani.labels import label_number
  import sys
  # Display system and polars versions
  print(f'My system version is {sys.version};\npolars version is {pl.__version__}')
My system version is 3.12.4 (main, Jul 1 2024, 00:48:18) [Clang 15.0.0 (clang-1500.3.9.4)];
polars version is 1.2.1
My system version is 3.12.4 (main, Jul 1 2024, 00:48:18) [Clang 15.0.0 (clang-1500.3.9.4)];
polars version is 1.2.1
  url ='https://raw.githubusercontent.com/AlexTheAnalyst/PandasYouTubeSeries/main/world_popu
  world_pop_raw = pl.read_csv(url)
  # Inspect output
  print(world_pop_raw)
shape: (234, 17)
        CCA3 Country
                                                                      Density (per km<sup>2</sup>)
 Rank
                                   Capital
                                                      ... Area (km²)
                                                                                           Grow
                                                          i64
                                                                       f64
 i64
        str
              \operatorname{str}
                                   str
                                                                                           f64
```

36	AFG	Afghanistan	Kabul	•••	652230	63.0587	1.02
138	ALB	Albania	Tirana	•••	28748	98.8702	0.99
34	DZA	Algeria	Algiers	•••	2381741	18.8531	1.01
213	ASM	American Samoa	Pago Pago		199	222.4774	0.98
203	AND	Andorra	Andorra la Vella		468	170.5641	1.01
•••	•••	•••	•••		•••	•••	
226	WLF	Wallis and Futuna	Mata-Utu	•••	142	81.493	0.99
172	ESH	Western Sahara	El Aaiún	•••	266000	2.1654	1.01
46	YEM	Yemen	Sanaa	•••	527968	63.8232	1.02
63	ZMB	Zambia	Lusaka	•••	752612	26.5976	1.02
74	ZWE	Zimbabwe	Harare		390757	41.7665	1.02
shape:	(201,						
shape: Rank	CCA3	Country	Capital		Area (km²)	v -	Grow
Rank 							
Rank		•	-	•••		v -	
Rank 	 str	 str					 f64
Rank i64			str		 i64	 f64	
Rank i64 36	str AFG	str Afghanistan	str Kabul		 i64 652230	f64 63.0587	 f64 1.02
Rank i64 36 138	str AFG ALB	str Afghanistan Albania	str Kabul Tirana Algiers	•••	 i64 652230 28748	f64 63.0587 98.8702	 f64 1.02 0.99
Rank i64 36 138 34	aFG ALB DZA	str Afghanistan Albania Algeria	str Kabul Tirana Algiers	•••	 i64 652230 28748 2381741	f64 63.0587 98.8702 18.8531	1.02 0.99
Rank i64 36 138 34 213	str AFG ALB DZA ASM	str Afghanistan Albania Algeria American Samoa	str Kabul Tirana Algiers Pago Pago		 i64 652230 28748 2381741 199	f64 63.0587 98.8702 18.8531 222.4774	 f64 1.02 0.99 1.01 0.98
Rank i64 36 138 34 213 203	afg ALB DZA ASM AND	str Afghanistan Albania Algeria American Samoa	str Kabul Tirana Algiers Pago Pago Andorra la Vella		 i64 652230 28748 2381741 199 468	f64 63.0587 98.8702 18.8531 222.4774 170.5641	 f64 1.02 0.99 1.01 0.98 1.01
Rank i64 36 138 34 213 203	str AFG ALB DZA ASM AND	str Afghanistan Albania Algeria American Samoa Andorra	str Kabul Tirana Algiers Pago Pago Andorra la Vella		 i64 652230 28748 2381741 199 468	f64 63.0587 98.8702 18.8531 222.4774 170.5641	1.02 0.99 1.01 0.98 1.01
Rank i64 36 138 34 213 203 226	str AFG ALB DZA ASM AND WLF	str Afghanistan Albania Algeria American Samoa Andorra Wallis and Futuna	str Kabul Tirana Algiers Pago Pago Andorra la Vella Mata-Utu		 i64 652230 28748 2381741 199 468 	f64 63.0587 98.8702 18.8531 222.4774 170.5641 81.493	1.02 0.99 1.01 0.98 1.01
Rank i64 36 138 34 213 203 226 172	str AFG ALB DZA ASM AND WLF	str Afghanistan Albania Algeria American Samoa Andorra Wallis and Futuna Western Sahara	str Kabul Tirana Algiers Pago Pago Andorra la Vella Mata-Utu El Aaiún		 i64 652230 28748 2381741 199 468 142 266000	f64 63.0587 98.8702 18.8531 222.4774 170.5641 81.493 2.1654	1.02 0.99 1.01 0.98 1.01 0.99 1.01
Rank i64 36 138 34 213 203 226 172 46	str AFG ALB DZA ASM AND WLF ESH YEM	str Afghanistan Albania Algeria American Samoa Andorra Wallis and Futuna Western Sahara Yemen	str Kabul Tirana Algiers Pago Pago Andorra la Vella Mata-Utu El Aaiún Sanaa		i64 652230 28748 2381741 199 468 142 266000 527968	f64 63.0587 98.8702 18.8531 222.4774 170.5641 81.493 2.1654 63.8232	1.02 0.99 1.01 0.98 1.01 0.99 1.01 1.02

world_pop_raw.dtypes

[Int64, String,

String,

String,

String,

Int64,

Int64,

Int64,

Int64,

Int64, Int64, Int64, Int64, Int64, Float64, Float64, Float64]

print(world_pop_raw.describe())

shape: (9, 18)

statistic	Rank	CCA3	Country		Area (km^2)	Density (per km ²)	Growth Ra
str	f64	str	str		f64	f64	f64
count	234.0	234	234	•••	232.0	230.0	232.0
null_count	0.0	0	0		2.0	4.0	2.0
mean	117.5	null	null		581663.74569	456.811652	1.009553
std	67.694165	null	null		1.7691e6	2083.740364	0.01339
min	1.0	ABW	Afghanistan	•••	1.0	0.0261	0.912
25%	59.0	null	null		2586.0	36.0935	1.002
50%	118.0	null	null		78865.0	96.7026	1.0079
75%	176.0	null	null		406752.0	236.9867	1.0165
max	234.0	ZWE	Zimbabwe		1.7098242e7	23172.2667	1.0691

shape: (9, 18)

statistic	Rank	CCA3	Country	•••	Area (km²)	Density (per km ²)	Growth Ra
str	f64	str	str		f64	f64	f64
count	234.0	234	234		232.0	230.0	232.0
null_count	0.0	0	0	•••	2.0	4.0	2.0
mean	117.5	null	null	•••	581663.74569	456.811652	1.009553
std	67.694165	null	null	•••	1.7691e6	2083.740364	0.01339
min	1.0	ABW	Afghanistan	•••	1.0	0.0261	0.912
25%	59.0	null	null	•••	2586.0	36.0935	1.002
50%	118.0	null	null	•••	78865.0	96.7026	1.0079
75%	176.0	null	null		406752.0	236.9867	1.0165

print(world_pop_raw.glimpse())

```
Rows: 234
Columns: 17
$ Rank
                                                                                                                        <i64> 36, 138, 34, 213, 203, 42, 224, 201, 33, 140
$ CCA3
                                                                                                                        <str> 'AFG', 'ALB', 'DZA', 'ASM', 'AND', 'AGO', 'AIA', 'ATG',
                                                                                                                        <str> 'Afghanistan', 'Albania', 'Algeria', 'American Samoa', '.
$ Country
$ Capital
                                                                                                                        <str> 'Kabul', 'Tirana', 'Algiers', 'Pago Pago', 'Andorra la Vonta de la 
                                                                                                                       <str> 'Asia', 'Europe', 'Africa', 'Oceania', 'Europe', 'Africa
$ Continent
$ 2022 Population
                                                                                                                       <:64> 41128771, 2842321, 44903225, 44273, 79824, 35588987, 158
                                                                                                                       <164> 38972230, 2866849, 43451666, 46189, 77700, 33428485, 155
$ 2020 Population
$ 2015 Population
                                                                                                                       <164> 33753499, 2882481, 39543154, 51368, 71746, 28127721, 145
                                                                                                                       <i64> 28189672, 2913399, 35856344, 54849, 71519, 23364185, 131
$ 2010 Population
                                                                                                                       <164> 19542982, 3182021, 30774621, 58230, 66097, 16394062, 110
$ 2000 Population
$ 1990 Population
                                                                                                                       <164> 10694796, 3295066, 25518074, 47818, 53569, 11828638, 831
$ 1980 Population
                                                                                                                       <i64> 12486631, 2941651, 18739378, 32886, 35611, 8330047, 6560
$ 1970 Population
                                                                                                                       <i64> 10752971, 2324731, 13795915, 27075, 19860, 6029700, 6283
$ Area (km<sup>2</sup>)
                                                                                                                        <164> 652230, 28748, 2381741, 199, 468, 1246700, 91, 442, 2780
                                                                                                                        <f64> 63.0587, 98.8702, 18.8531, 222.4774, 170.5641, 28.5466,
$ Density (per km<sup>2</sup>)
                                                                                                                        <f64> 1.0257, 0.9957, 1.0164, 0.9831, 1.01, 1.0315, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.0066, 1.
$ Growth Rate
$ World Population Percentage <f64> 0.52, 0.04, 0.56, 0.0, 0.0, 0.45, 0.0, 0.0, 0.57, 0.03
```

1.0691

None

\$ 1980 Population \$ 1970 Population

Rows: 234 Columns: 17 <i64> 36, 138, 34, 213, 203, 42, 224, 201, 33, 140 \$ Rank <str> 'AFG', 'ALB', 'DZA', 'ASM', 'AND', 'AGO', 'AIA', 'ATG', \$ CCA3 <str> 'Afghanistan', 'Albania', 'Algeria', 'American Samoa', '. \$ Country <str> 'Kabul', 'Tirana', 'Algiers', 'Pago Pago', 'Andorra la Vonta de la \$ Capital <str> 'Asia', 'Europe', 'Africa', 'Oceania', 'Europe', 'Africa \$ Continent \$ 2022 Population <:64> 41128771, 2842321, 44903225, 44273, 79824, 35588987, 158 \$ 2020 Population <164> 38972230, 2866849, 43451666, 46189, 77700, 33428485, 155 <164> 33753499, 2882481, 39543154, 51368, 71746, 28127721, 1453 \$ 2015 Population <i64> 28189672, 2913399, 35856344, 54849, 71519, 23364185, 131 \$ 2010 Population <164> 19542982, 3182021, 30774621, 58230, 66097, 16394062, 1104 \$ 2000 Population \$ 1990 Population <164> 10694796, 3295066, 25518074, 47818, 53569, 11828638, 831

<i64> 12486631, 2941651, 18739378, 32886, 35611, 8330047, 6560

<i64> 10752971, 2324731, 13795915, 27075, 19860, 6029700, 6283

```
$ Area (km<sup>2</sup>)
                                 <164> 652230, 28748, 2381741, 199, 468, 1246700, 91, 442, 2780
$ Density (per km²)
                                 <f64> 63.0587, 98.8702, 18.8531, 222.4774, 170.5641, 28.5466,
                                 <f64> 1.0257, 0.9957, 1.0164, 0.9831, 1.01, 1.0315, 1.0066, 1.0066
$ Growth Rate
$ World Population Percentage <f64> 0.52, 0.04, 0.56, 0.0, 0.0, 0.45, 0.0, 0.57, 0.03
None
  print(world_pop_raw.null_count())
shape: (1, 17)
         CCA3
                                                      Density (per km<sup>2</sup>)
                                                                                           World Popu
 Rank
                Country
                          Capital
                                        Area (km²)
                                                                            Growth Rate
 u32
         u32
                u32
                          u32
                                         u32
                                                       u32
                                                                            u32
                                                                                           u32
                0
 0
         0
                          0
                                         2
                                                                            2
                                                                                           0
                                                       4
shape: (1, 17)
                          Capital
 Rank
         CCA3
                                         Area (km<sup>2</sup>)
                                                       Density (per km<sup>2</sup>)
                                                                            Growth Rate
                                                                                           World Popu
                Country
                                                                                            ---
 u32
         u32
                u32
                          u32
                                         u32
                                                       u32
                                                                            u32
                                                                                           u32
 0
         0
                0
                          0
                                         2
                                                       4
                                                                            2
                                                                                           0
  print(
       world_pop_raw
       .select(cs.all().is_null().sum())
       .glimpse()
   )
```

Rows: 1

Columns: 17

\$ Rank	<u32></u32>	0
\$ CCA3	<u32></u32>	0
\$ Country	<u32></u32>	0
\$ Capital	<u32></u32>	0
\$ Continent	<u32></u32>	0
\$ 2022 Population	<u32></u32>	4

```
<u32> 1
$ 2020 Population
$ 2015 Population
                            <u32> 4
                            <u32> 7
$ 2010 Population
$ 2000 Population
                            <u32> 7
                            <u32> 5
$ 1990 Population
                            <u32> 5
$ 1980 Population
$ 1970 Population
                            <u32> 4
                            <u32> 2
$ Area (km<sup>2</sup>)
$ Density (per km<sup>2</sup>)
                            <u32> 4
                             <u32> 2
$ Growth Rate
$ World Population Percentage <u32> 0
None
Rows: 1
Columns: 17
$ Rank
                            <u32> 0
$ CCA3
                             <u32> 0
                            <u32> 0
$ Country
                            <u32> 0
$ Capital
                            <u32> 0
$ Continent
$ 2022 Population
                            <u32> 4
                            <u32> 1
$ 2020 Population
                            <u32> 4
$ 2015 Population
                            <u32> 7
$ 2010 Population
                            <u32> 7
$ 2000 Population
                            <u32> 5
$ 1990 Population
                            <u32> 5
$ 1980 Population
$ 1970 Population
                            <u32> 4
                            <u32> 2
$ Area (km<sup>2</sup>)
$ Density (per km<sup>2</sup>)
                            <u32> 4
                             <u32> 2
$ Growth Rate
$ World Population Percentage <u32> 0
None
  # Unique column values
  print(
      world_pop_raw
```

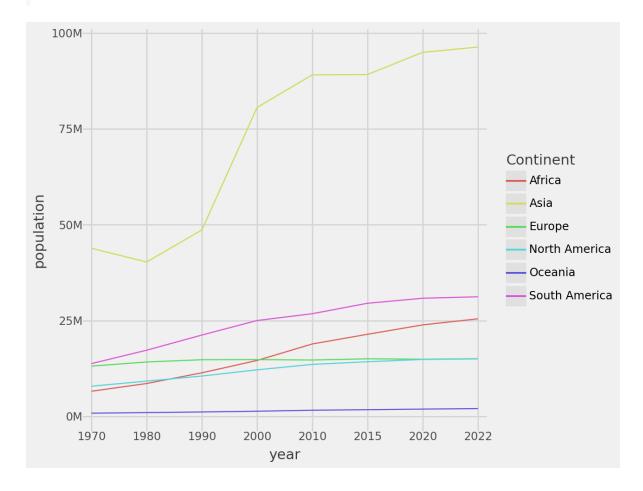
.unique(subset=['CCA3', 'Country'], maintain_order=True)

.get_column('Country')

)

```
shape: (234,)
Series: 'Country' [str]
    "Afghanistan"
    "Albania"
    "Algeria"
    "American Samoa"
    "Andorra"
    "Wallis and Futuna"
    "Western Sahara"
    "Yemen"
    "Zambia"
    "Zimbabwe"
shape: (234,)
Series: 'Country' [str]
    "Afghanistan"
    "Albania"
    "Algeria"
    "American Samoa"
    "Andorra"
    "Wallis and Futuna"
    "Western Sahara"
    "Yemen"
    "Zambia"
    "Zimbabwe"
]
  continent_mean_pop = (
      world_pop_raw
      .group_by('Continent')
       .agg(cs.ends_with('Population').mean())
       .sort('2022 Population', descending=False)
       .unpivot(cs.numeric(), index='Continent', variable_name='year', value_name='population
      .with_columns(year=pl.col('year').str.strip_chars(' Population'))
  )
```

```
ggplot(continent_mean_pop, aes('year', 'population', group='Continent'))
+ geom_line(aes(color='Continent'))
+ scale_y_continuous(
    labels=label_number(scale=1e-6, suffix='M'),
    expand=(0.05, 0.02)
)
+ scale_x_discrete(expand=(0.02, 0.02))
+ theme_538()
)
```



```
(
    ggplot(continent_mean_pop, aes('year', 'population', group='Continent'))
    + geom_line(aes(color='Continent'))
    + scale_y_continuous(
```

```
labels=label_number(scale=1e-6, suffix='M'),
    expand=(0.05, 0.02)
)

+ scale_x_discrete(expand=(0.02, 0.02))
+ scale_color_manual(values=['#9fa19c', '#92b854', '#9fa19c', '#9fa19c', '#9fa19c', '#
+ guides(shape=None, color=None, fill=None)
+ theme_538()
+ labs(
    x=None,
    title='Average Population by Continent Over the Years',
    caption='Data Source:\nhttps://raw.githubusercontent.com/AlexTheAnalyst/PandasYouT
)
+ theme(
    plot_title=element_text(ha=0, margin={'t': 15, 'b': 15}),
    legend_position='bottom',
    axis_title_x=element_blank()
)
)
```

