# Exploratory Data Analysis with polars Library

Alier Reng

```python
# 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
```

```python
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)

 Rank   CCA3   Country              Capital             …  Area (km²)   Density (per km²)   Grow
 ---    ---    ---                  ---                    ---          ---                 ---
 i64    str    str                  str                    i64          f64                 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: (234, 17)

 Rank   CCA3   Country             Capital            …   Area (km²)   Density (per km²)    Grow
 ---    ---    ---                 ---                    ---          ---                  ---
 i64    str    str                 str                    i64          f64                  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
```

```
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 (km²)    Density (per km²)  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 km²)  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
```

```
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', 
$ Country                     <str> 'Afghanistan', 'Albania', 'Algeria', 'American Samoa', '
$ Capital                     <str> 'Kabul', 'Tirana', 'Algiers', 'Pago Pago', 'Andorra la Ve
$ Continent                   <str> 'Asia', 'Europe', 'Africa', 'Oceania', 'Europe', 'Africa
$ 2022 Population             <i64> 41128771, 2842321, 44903225, 44273, 79824, 35588987, 1585
$ 2020 Population             <i64> 38972230, 2866849, 43451666, 46189, 77700, 33428485, 1558
$ 2015 Population             <i64> 33753499, 2882481, 39543154, 51368, 71746, 28127721, 1452
$ 2010 Population             <i64> 28189672, 2913399, 35856344, 54849, 71519, 23364185, 1317
$ 2000 Population             <i64> 19542982, 3182021, 30774621, 58230, 66097, 16394062, 1104
$ 1990 Population             <i64> 10694796, 3295066, 25518074, 47818, 53569, 11828638, 8316
$ 1980 Population             <i64> 12486631, 2941651, 18739378, 32886, 35611, 8330047, 6560
$ 1970 Population             <i64> 10752971, 2324731, 13795915, 27075, 19860, 6029700, 6283
$ Area (km²)                  <i64> 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, 
$ Growth Rate                 <f64> 1.0257, 0.9957, 1.0164, 0.9831, 1.01, 1.0315, 1.0066, 1.0
$ World Population Percentage <f64> 0.52, 0.04, 0.56, 0.0, 0.0, 0.45, 0.0, 0.0, 0.57, 0.03

None
```

```
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', 
$ Country                     <str> 'Afghanistan', 'Albania', 'Algeria', 'American Samoa', '
$ Capital                     <str> 'Kabul', 'Tirana', 'Algiers', 'Pago Pago', 'Andorra la Ve
$ Continent                   <str> 'Asia', 'Europe', 'Africa', 'Oceania', 'Europe', 'Africa
$ 2022 Population             <i64> 41128771, 2842321, 44903225, 44273, 79824, 35588987, 1585
$ 2020 Population             <i64> 38972230, 2866849, 43451666, 46189, 77700, 33428485, 1558
$ 2015 Population             <i64> 33753499, 2882481, 39543154, 51368, 71746, 28127721, 1452
$ 2010 Population             <i64> 28189672, 2913399, 35856344, 54849, 71519, 23364185, 1317
$ 2000 Population             <i64> 19542982, 3182021, 30774621, 58230, 66097, 16394062, 1104
$ 1990 Population             <i64> 10694796, 3295066, 25518074, 47818, 53569, 11828638, 8316
$ 1980 Population             <i64> 12486631, 2941651, 18739378, 32886, 35611, 8330047, 6560
$ 1970 Population             <i64> 10752971, 2324731, 13795915, 27075, 19860, 6029700, 6283
$ Area (km²)                 <i64> 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,
$ Growth Rate                <f64> 1.0257, 0.9957, 1.0164, 0.9831, 1.01, 1.0315, 1.0066, 1.0
$ World Population Percentage <f64> 0.52, 0.04, 0.56, 0.0, 0.0, 0.45, 0.0, 0.0, 0.57, 0.03

None
```

```
print(world_pop_raw.null_count())
```

```
shape: (1, 17)

 Rank   CCA3   Country   Capital   …   Area (km²)   Density (per km²)   Growth Rate   World Popu
 ---    ---    ---       ---           ---          ---                 ---           ---
 u32    u32    u32       u32           u32          u32                 u32           u32

 0      0      0         0         …   2            4                   2             0
```

```
shape: (1, 17)

 Rank   CCA3   Country   Capital   …   Area (km²)   Density (per km²)   Growth Rate   World Popu
 ---    ---    ---       ---           ---          ---                 ---           ---
 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()
)
```

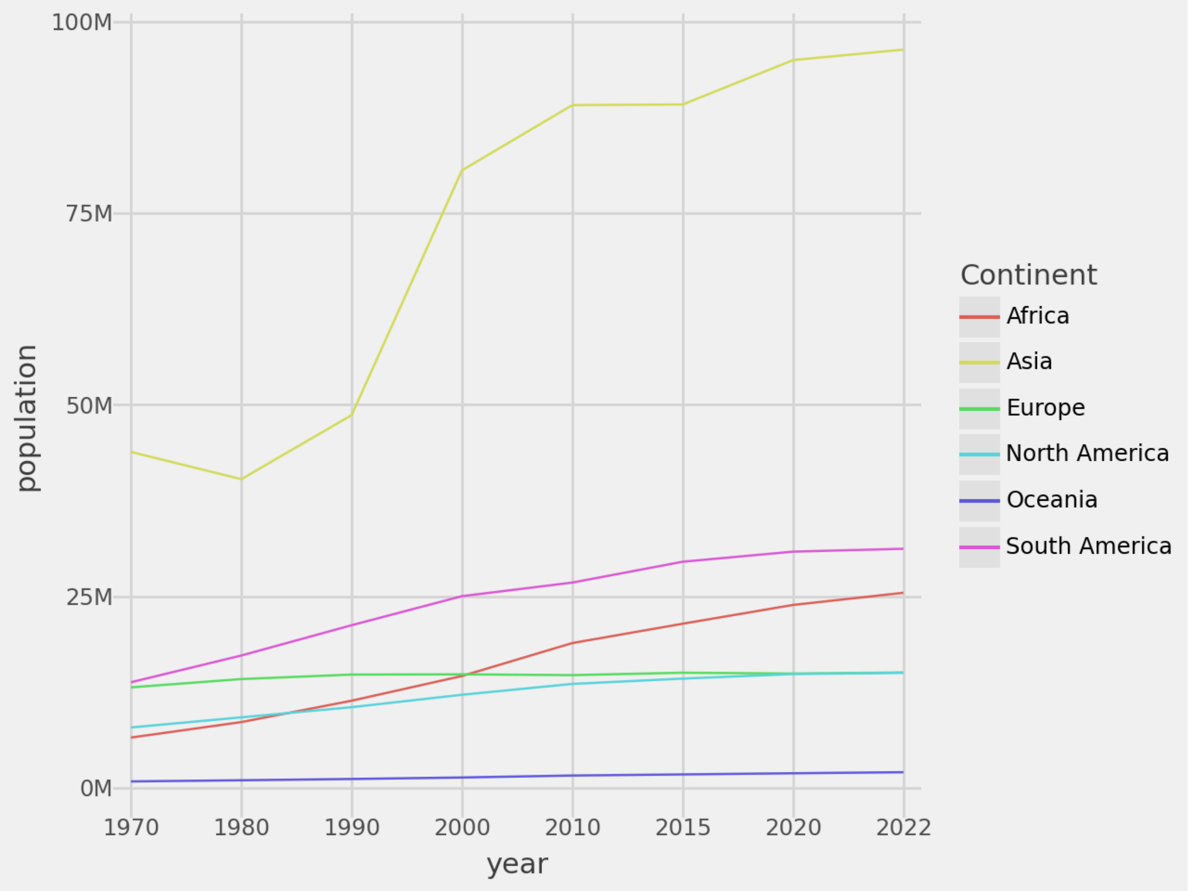
```
Rows: 1
Columns: 17
$ Rank                       <u32> 0
$ CCA3                       <u32> 0
$ Country                    <u32> 0
$ Capital                    <u32> 0
$ Continent                  <u32> 0
$ 2022 Population            <u32> 4
$ 2020 Population            <u32> 1
$ 2015 Population            <u32> 4
$ 2010 Population            <u32> 7
$ 2000 Population            <u32> 7
$ 1990 Population            <u32> 5
$ 1980 Population            <u32> 5
$ 1970 Population            <u32> 4
$ Area (km²)                 <u32> 2
$ Density (per km²)          <u32> 4
$ Growth Rate                <u32> 2
$ World Population Percentage <u32> 0

None
Rows: 1
Columns: 17
$ Rank                       <u32> 0
$ CCA3                       <u32> 0
$ Country                    <u32> 0
$ Capital                    <u32> 0
$ Continent                  <u32> 0
$ 2022 Population            <u32> 4
$ 2020 Population            <u32> 1
$ 2015 Population            <u32> 4
$ 2010 Population            <u32> 7
$ 2000 Population            <u32> 7
$ 1990 Population            <u32> 5
$ 1980 Population            <u32> 5
$ 1970 Population            <u32> 4
$ Area (km²)                 <u32> 2
$ Density (per km²)          <u32> 4
$ Growth Rate                <u32> 2
$ World Population Percentage <u32> 0

None
```

```python
# Unique column values
print(
    world_pop_raw
    .unique(subset=['CCA3', 'Country'], maintain_order=True)
    .get_column('Country')
)
```

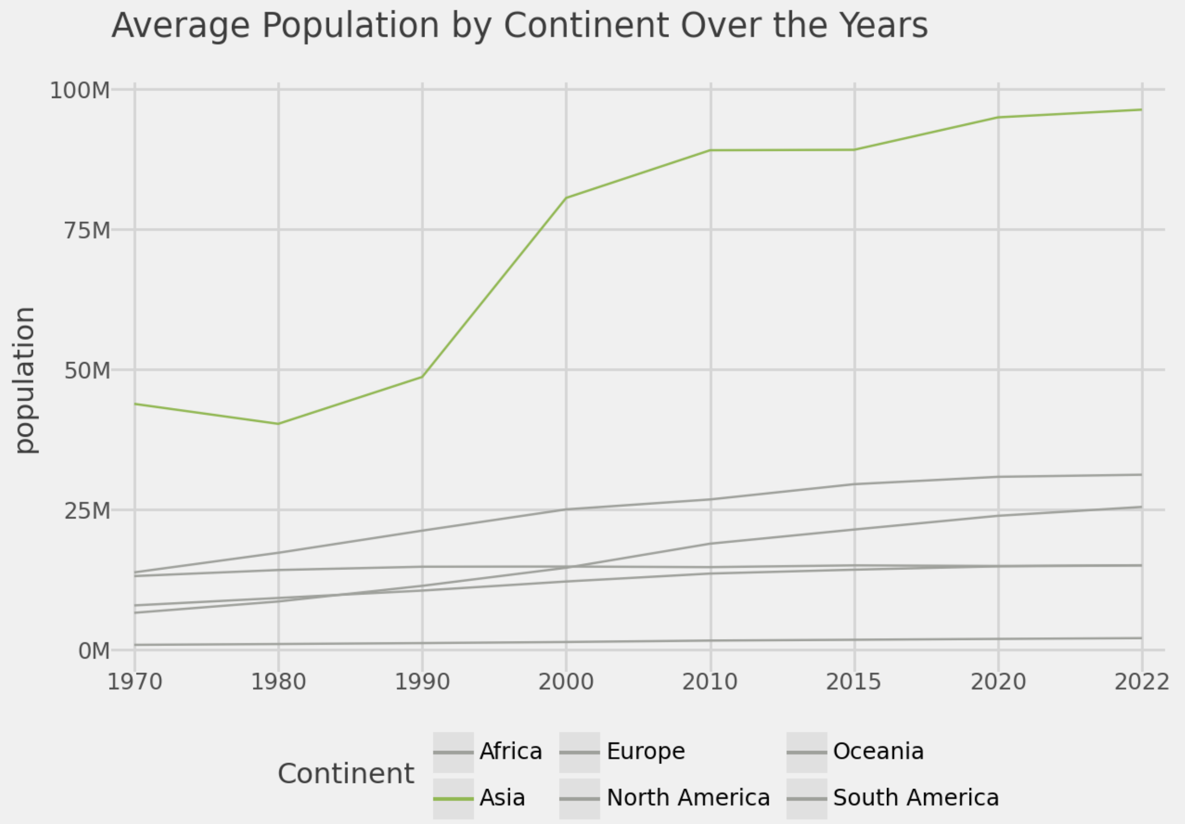
```
shape: (234,)
Series: 'Country' [str]
[
	"Afghanistan"
	"Albania"
	"Algeria"
	"American Samoa"
	"Andorra"
	…
	"Wallis and Futuna"
	"Western Sahara"
	"Yemen"
	"Zambia"
	"Zimbabwe"
]
shape: (234,)
Series: 'Country' [str]
[
	"Afghanistan"
	"Albania"
	"Algeria"
	"American Samoa"
	"Andorra"
	…
	"Wallis and Futuna"
	"Western Sahara"
	"Yemen"
	"Zambia"
	"Zimbabwe"
]
```

```python
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()
    )
)
```



Data Source:
https://raw.githubusercontent.com/AlexTheAnalyst/PandasYouTubeSeries/main/world_population.csv