In [1]:
```
# import libraries
import pandas as pd # read in csv
import matplotlib.pyplot as plt # plotting
import seaborn as sns # plotting, but makes things easier
```

In [2]:
```
# read in data from csv file
myDataFromFile = pd.read_csv("../datasets/006DataFile.csv")
```

In the line of code above, what is the:

- name of the library used to load the file? pandas
- name of the pandas function we use to read the data file? read_csv()
- data file name? "006DataFile.csv"
- name of the variable used to store the file? myDataFromFile
- name of the folder containing the data file? "datasets"

In [3]:
```
# show entire pandas
display(myDataFromFile)
```

	VarA	VarB
0	0.979109	-0.128890
1	0.196564	0.403177
2	0.260841	0.682448
3	2.432641	-0.295968
4	-0.689790	-0.088941
...	...	...
95	2.416932	-1.065406
96	4.166266	-1.053911
97	-0.203719	0.610032
98	1.232813	-0.744738
99	0.833993	-0.372451

100 rows × 2 columns

What are the dimensions (the size) of the data? 100 rows x 2 columns (100 observations, 2 variables)

In [4]:
```
# Extra tip: read in data from clipboard
cb = pd.read_clipboard()
cb
```

Out[4]:

	Total population	Population aged 0–14 (%)	Population aged 15–64 (%)	Population aged 65+ (%)
1950	4 284 000	40.7	57.3	2.0
1955	4 517 000	41.0	56.9	2.2
1960	4 829 000	41.3	56.3	2.3
1965	5 175 000	42.2	55.2	2.5
1970	5 625 000	43.3	53.9	2.8
1975	6 155 000	44.2	52.8	3.0
1980	6 823 000	45.6	51.2	3.2
1985	7 728 000	46.7	50.0	3.3
1990	8 811 000	47.3	49.5	3.3
1995	10 090 000	47.1	49.8	3.1
2000	11 608 000	46.8	50.5	2.8
2005	13 422 000	46.5	50.9	2.6
2010	15 605 000	46.2	51.3	2.5
2015	18 111 000	45.6	52.0	2.4
2020	20 903 000	44.4	53.2	2.4

- What are the names of the columns of the data copied from Wikipedia? Total population, population aged 0-14 (%), population aged 15-64 (%), population aged 65+ (%)
- What is the name of the variable I used to save the data in Python? "cb"

In [5]:
```
# display top 5 rows
myDataFromFile.head()
```

Out[5]:

	VarA	VarB
0	0.979109	-0.128890
1	0.196564	0.403177
2	0.260841	0.682448
3	2.432641	-0.295968
4	-0.689790	-0.088941

In [6]:
```
# display bottom 5 rows
myDataFromFile.tail()
```

Out[6]:

	VarA	VarB
95	2.416932	-1.065406
96	4.166266	-1.053911
97	-0.203719	0.610032
98	1.232813	-0.744738
99	0.833993	-0.372451

In [7]:
```
# show all methods that is recognized by a variable
dir(myDataFromFile)
```

Out[7]:
```
['T',
 'VarA',
 'VarB',
 '_AXIS_LEN',
 '_AXIS_ORDERS',
 '_AXIS_TO_AXIS_NUMBER',
 '_HANDLED_TYPES',
 '__abs__',
 '__add__',
 '__and__',
 '__annotations__',
 '__array__',
 '__array_priority__',
 '__array_ufunc__',
 '__array_wrap__',
 '__bool__',
 '__class__',
 '__contains__',
 '__copy__',
 '__deepcopy__',
 '__delattr__',
 '__delitem__',
 '__dict__',
 '__dir__',
 '__divmod__',
 '__doc__',
 '__eq__',
 '__finalize__',
 '__floordiv__',
 '__format__',
 '__ge__',
 '__getattr__',
 '__getattribute__',
 '__getitem__',
 '__getstate__',
 '__gt__',
 '__hash__',
 '__iadd__',
 '__iand__',
 '__ifloordiv__',
 '__imod__',
 '__imul__',
 '__init__',
 '__init_subclass__',
 '__invert__',
 '__ior__',
 '__ipow__',
 '__isub__',
 '__iter__',
 '__itruediv__',
 '__ixor__',
 '__le__',
 '__len__',
 '__lt__',
 '__matmul__',
 '__mod__',
 '__module__',
 '__mul__',
 '__ne__',
 '__neg__',
```

```
'__new__',
'__nonzero__',
'__or__',
'__pos__',
'__pow__',
'__radd__',
'__rand__',
'__rdivmod__',
'__reduce__',
'__reduce_ex__',
'__repr__',
'__rfloordiv__',
'__rmatmul__',
'__rmod__',
'__rmul__',
'__ror__',
'__round__',
'__rpow__',
'__rsub__',
'__rtruediv__',
'__rxor__',
'__setattr__',
'__setitem__',
'__setstate__',
'__sizeof__',
'__str__',
'__sub__',
'__subclasshook__',
'__truediv__',
'__weakref__',
'__xor__',
'_accessors',
'_accum_func',
'_add_numeric_operations',
'_agg_by_level',
'_agg_examples_doc',
'_agg_summary_and_see_also_doc',
'_align_frame',
'_align_series',
'_append',
'_arith_method',
'_as_manager',
'_attrs',
'_box_col_values',
'_can_fast_transpose',
'_check_inplace_and_allows_duplicate_labels',
'_check_inplace_setting',
'_check_is_chained_assignment_possible',
'_check_label_or_level_ambiguity',
'_check_setitem_copy',
'_clear_item_cache',
'_clip_with_one_bound',
'_clip_with_scalar',
'_cmp_method',
'_combine_frame',
'_consolidate',
'_consolidate_inplace',
'_construct_axes_dict',
'_construct_axes_from_arguments',
'_construct_result',
```

```
 '_constructor',
 '_constructor_sliced',
 '_convert',
 '_count_level',
 '_data',
 '_dir_additions',
 '_dir_deletions',
 '_dispatch_frame_op',
 '_drop_axis',
 '_drop_labels_or_levels',
 '_ensure_valid_index',
 '_find_valid_index',
 '_flags',
 '_from_arrays',
 '_from_mgr',
 '_get_agg_axis',
 '_get_axis',
 '_get_axis_name',
 '_get_axis_number',
 '_get_axis_resolvers',
 '_get_block_manager_axis',
 '_get_bool_data',
 '_get_cleaned_column_resolvers',
 '_get_column_array',
 '_get_index_resolvers',
 '_get_item_cache',
 '_get_label_or_level_values',
 '_get_numeric_data',
 '_get_value',
 '_getitem_bool_array',
 '_getitem_multilevel',
 '_gotitem',
 '_hidden_attrs',
 '_indexed_same',
 '_info_axis',
 '_info_axis_name',
 '_info_axis_number',
 '_info_repr',
 '_init_mgr',
 '_inplace_method',
 '_internal_names',
 '_internal_names_set',
 '_is_copy',
 '_is_homogeneous_type',
 '_is_label_or_level_reference',
 '_is_label_reference',
 '_is_level_reference',
 '_is_mixed_type',
 '_is_view',
 '_iset_item',
 '_iset_item_mgr',
 '_iset_not_inplace',
 '_item_cache',
 '_iter_column_arrays',
 '_ixs',
 '_join_compat',
 '_logical_func',
 '_logical_method',
 '_maybe_cache_changed',
 '_maybe_update_cacher',
```

```
'_metadata',
'_mgr',
'_min_count_stat_function',
'_needs_reindex_multi',
'_protect_consolidate',
'_reduce',
'_reduce_axis1',
'_reindex_axes',
'_reindex_columns',
'_reindex_index',
'_reindex_multi',
'_reindex_with_indexers',
'_rename',
'_replace_columnwise',
'_repr_data_resource_',
'_repr_fits_horizontal_',
'_repr_fits_vertical_',
'_repr_html_',
'_repr_latex_',
'_reset_cache',
'_reset_cacher',
'_sanitize_column',
'_series',
'_set_axis',
'_set_axis_name',
'_set_axis_nocheck',
'_set_is_copy',
'_set_item',
'_set_item_frame_value',
'_set_item_mgr',
'_set_value',
'_setitem_array',
'_setitem_frame',
'_setitem_slice',
'_slice',
'_stat_axis',
'_stat_axis_name',
'_stat_axis_number',
'_stat_function',
'_stat_function_ddof',
'_take_with_is_copy',
'_to_dict_of_blocks',
'_typ',
'_update_inplace',
'_validate_dtype',
'_values',
'_where',
'abs',
'add',
'add_prefix',
'add_suffix',
'agg',
'aggregate',
'align',
'all',
'any',
'append',
'apply',
'applymap',
'asfreq',
```

```
 'asof',
 'assign',
 'astype',
 'at',
 'at_time',
 'attrs',
 'axes',
 'backfill',
 'between_time',
 'bfill',
 'bool',
 'boxplot',
 'clip',
 'columns',
 'combine',
 'combine_first',
 'compare',
 'convert_dtypes',
 'copy',
 'corr',
 'corrwith',
 'count',
 'cov',
 'cummax',
 'cummin',
 'cumprod',
 'cumsum',
 'describe',
 'diff',
 'div',
 'divide',
 'dot',
 'drop',
 'drop_duplicates',
 'droplevel',
 'dropna',
 'dtypes',
 'duplicated',
 'empty',
 'eq',
 'equals',
 'eval',
 'ewm',
 'expanding',
 'explode',
 'ffill',
 'fillna',
 'filter',
 'first',
 'first_valid_index',
 'flags',
 'floordiv',
 'from_dict',
 'from_records',
 'ge',
 'get',
 'groupby',
 'gt',
 'head',
 'hist',
```

```
 'iat',
 'idxmax',
 'idxmin',
 'iloc',
 'index',
 'infer_objects',
 'info',
 'insert',
 'interpolate',
 'isin',
 'isna',
 'isnull',
 'items',
 'iteritems',
 'iterrows',
 'itertuples',
 'join',
 'keys',
 'kurt',
 'kurtosis',
 'last',
 'last_valid_index',
 'le',
 'loc',
 'lookup',
 'lt',
 'mad',
 'mask',
 'max',
 'mean',
 'median',
 'melt',
 'memory_usage',
 'merge',
 'min',
 'mod',
 'mode',
 'mul',
 'multiply',
 'ndim',
 'ne',
 'nlargest',
 'notna',
 'notnull',
 'nsmallest',
 'nunique',
 'pad',
 'pct_change',
 'pipe',
 'pivot',
 'pivot_table',
 'plot',
 'pop',
 'pow',
 'prod',
 'product',
 'quantile',
 'query',
 'radd',
 'rank',
```

```
 'rdiv',
 'reindex',
 'reindex_like',
 'rename',
 'rename_axis',
 'reorder_levels',
 'replace',
 'resample',
 'reset_index',
 'rfloordiv',
 'rmod',
 'rmul',
 'rolling',
 'round',
 'rpow',
 'rsub',
 'rtruediv',
 'sample',
 'select_dtypes',
 'sem',
 'set_axis',
 'set_flags',
 'set_index',
 'shape',
 'shift',
 'size',
 'skew',
 'slice_shift',
 'sort_index',
 'sort_values',
 'squeeze',
 'stack',
 'std',
 'style',
 'sub',
 'subtract',
 'sum',
 'swapaxes',
 'swaplevel',
 'tail',
 'take',
 'to_clipboard',
 'to_csv',
 'to_dict',
 'to_excel',
 'to_feather',
 'to_gbq',
 'to_hdf',
 'to_html',
 'to_json',
 'to_latex',
 'to_markdown',
 'to_numpy',
 'to_parquet',
 'to_period',
 'to_pickle',
 'to_records',
 'to_sql',
 'to_stata',
 'to_string',
```

```
'to_timestamp',
'to_xarray',
'to_xml',
'transform',
'transpose',
'truediv',
'truncate',
'tz_convert',
'tz_localize',
'unstack',
'update',
'value_counts',
'values',
'var',
'where',
'xs']
```

In [8]:
```
# output summary of dataframe
myDataFromFile.describe()
```

Out[8]:

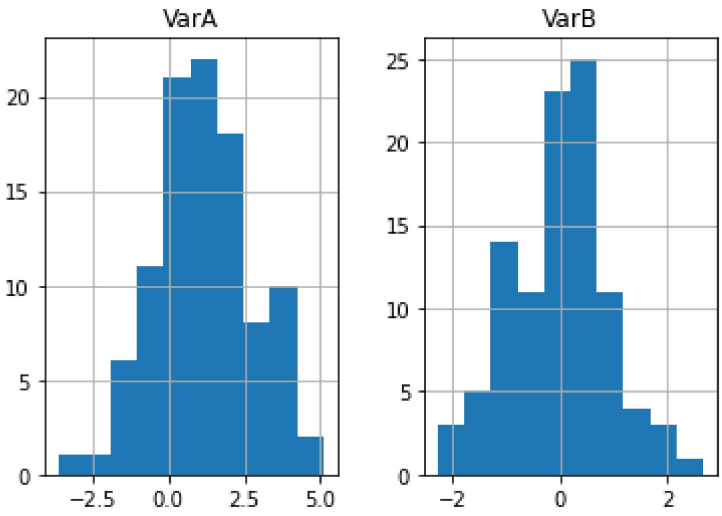
	VarA	VarB
count	100.000000	100.000000
mean	1.195657	0.017781
std	1.620649	0.924191
min	-3.640916	-2.275922
25%	0.247035	-0.653144
50%	1.084324	0.049298
75%	2.338135	0.615286
max	5.092458	2.669149

Describe in your own words what the method describe returns of the data. Describe the measures the method returns.

describe() method returns a simple summary of a dataframe (e.g. the values you may need to create a simple histogram or boxplot). Specifically, it returns the number of data points, mean, standard deviation, min, max, and the quartiles

In [9]:
```
# plot simple histogram
myDataFromFile.hist()
```

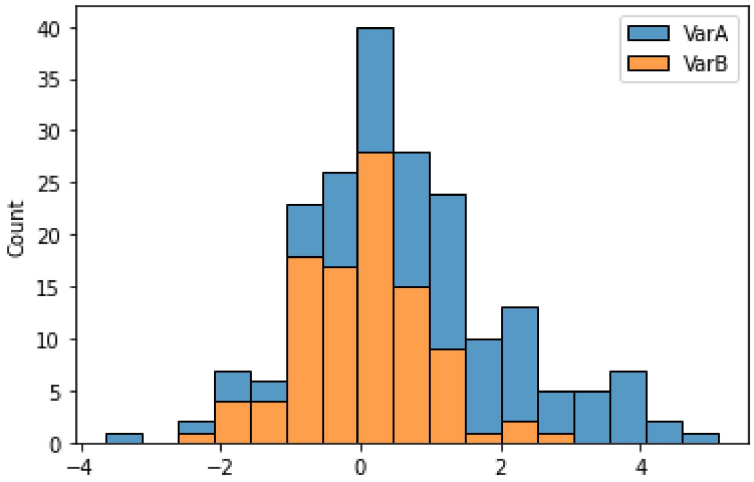
Out[9]:
```
array([[<AxesSubplot:title={'center':'VarA'}>,
        <AxesSubplot:title={'center':'VarB'}>]], dtype=object)
```



- What are the titles of the two histograms in the figures created by the method .hist? "VarA" and "VarB"
- How do the titles of the histograms relate to the data? One of the histogram plots the frequency of VarA (titled VarA) and the other graph plots the frequency of VarB (titled VarB).
- What are the values in the x-axis of the histograms? Numerical values of VarA and VarB (binned)

In [10]:
```
# plot histogram with seaborn
sns.histplot(myDataFromFile, multiple="stack")
```

Out[10]:
```
<AxesSubplot:ylabel='Count'>
```



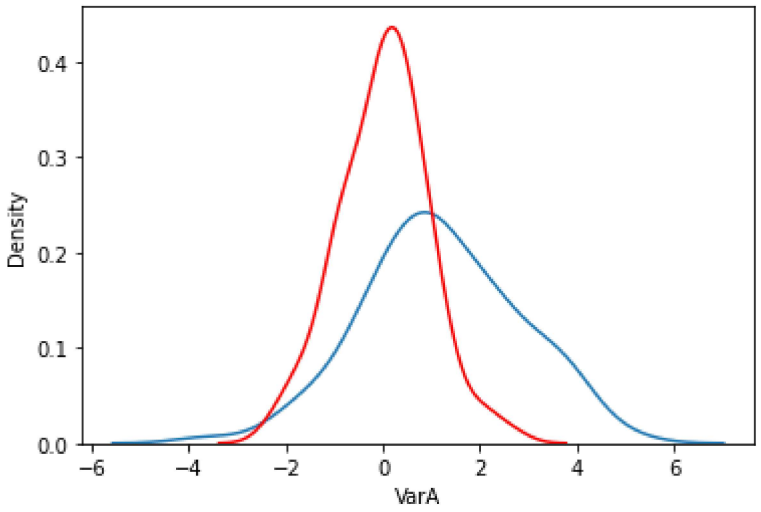
In [11]:
```
# show arguments for a method
sns.histplot?
```

report three arguments of .hist:

1. hue
2. weights
3. stat

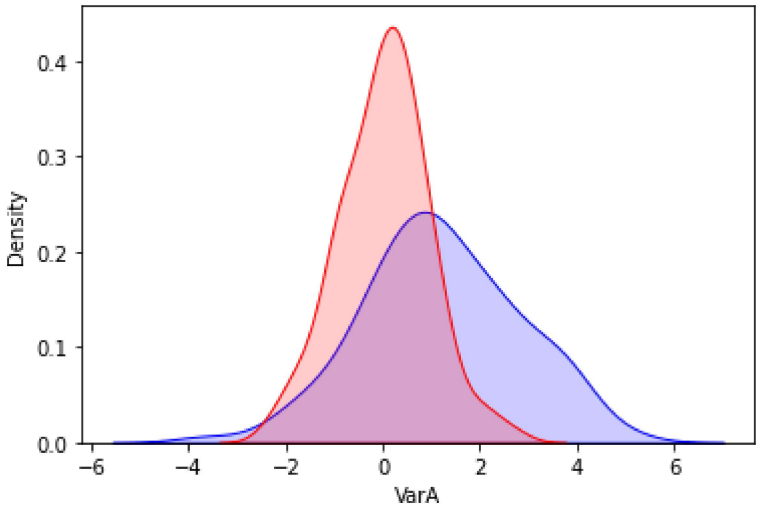
In [12]:
```
# plot KDE: plot each variable in turn, and use 'color' in second to set curves apart
sns.kdeplot(myDataFromFile["VarA"])
sns.kdeplot(myDataFromFile["VarB"], color="r")
```

Out[12]:
```
<AxesSubplot:xlabel='VarA', ylabel='Density'>
```



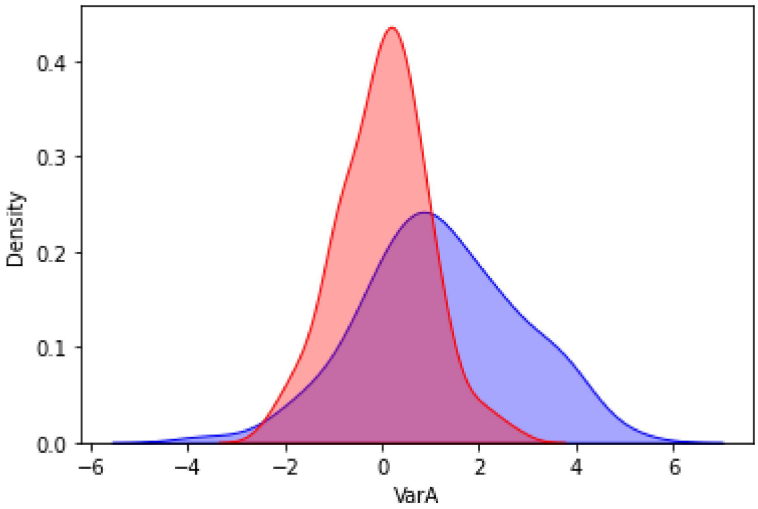
In [13]:
```
# plot KDE with extra arguments
sns.kdeplot(myDataFromFile["VarA"], color="b", fill=True, alpha=0.2)
sns.kdeplot(myDataFromFile["VarB"], color="r", fill=True, alpha=0.2)
```

Out[13]:
```
<AxesSubplot:xlabel='VarA', ylabel='Density'>
```



In [14]:
```
# plot KDE with adjusted alpha
sns.kdeplot(myDataFromFile["VarA"], color="b", fill=True, alpha=0.35)
sns.kdeplot(myDataFromFile["VarB"], color="r", fill=True, alpha=0.35)
```

Out[14]:
```
<AxesSubplot:xlabel='VarA', ylabel='Density'>
```



Make the same plot as above but increaset the transparency of the histograms to 0.35. How does it look? The colored portions of the graph look darker (i.e. less transparent) when alpha is set to 0.35 compared to when it's set to 0.2.

In [15]:
```
# create dataframe to write out
mySummary = myDataFromFile.describe()
print(mySummary)
```

```
              VarA        VarB
count  100.000000  100.000000
mean     1.195657    0.017781
std      1.620649    0.924191
min     -3.640916   -2.275922
25%      0.247035   -0.653144
50%      1.084324    0.049298
75%      2.338135    0.615286
max      5.092458    2.669149
```

What is the name of the variable we saved the output of the method .describe() in?

"mySummary"

In [16]:
```
# output dataframe to csv file
mySummary.to_csv("mySummary.csv")
```

In [17]:
```
# read in csv file to check that output worked correctly
mySumFF = pd.read_csv("mySummary.csv")
display(mySumFF)
```

	Unnamed: 0	VarA	VarB
0	count	100.000000	100.000000
1	mean	1.195657	0.017781
2	std	1.620649	0.924191
3	min	-3.640916	-2.275922
4	25%	0.247035	-0.653144
5	50%	1.084324	0.049298
6	75%	2.338135	0.615286
7	max	5.092458	2.669149