Tolga Keskinoglu
Matriculation #: 23188690

# Question

Can Franconia's wine region now include Bamberg due to warmer weather?

Despite the negative effects of climate change, some wine varietals thrive in warmer climates. For instance, Bamberg, once a wine-producing region, could potentially regain its status due to rising temperatures. We compare the climate of nearby Würzburg, at the heart of the Franconian winemaking region, to Bamberg's which is currently just outside of the Franconian winemaking region.

# Data Sources

Deutscher Wetterdienst (DWD) because it is well-maintained and has many direct measures of weather and contains the following data:

- QN_4 - quality level of the data in the following columns
- MO_N - monthly mean of cloud cover
- MO_TT - monthly mean of daily temperature means in 2m height
- MO_TX - monthly mean of daily temperature maxima in 2m height
- MO_TN - monthly mean of daily temperature minima in 2m height
- MO_FK - monthly mean of daily wind speed
- MX_TX - monthly maximum of daily temperature maxima in 2m height
- MX_FX - monthly maximum of daily wind speed
- MX_TN - monthly minimum of daily temperature minima in 2m height
- MO_SD_S - monthly sum of sunshine duration
- QN_6 - quality level of the data in the following columns
- MO_RR - monthly sum of precipitation height
- MX_RS - monthly maximum of daily precipitation height

```
STATIONS_ID;MESS_DATUM_BEGINN;MESS_DATUM_ENDE;QN_4;MO_N;MO_TT;MO_TX;MO_TN;MO_FK;MX_TX;MX_FX;MX_TN;MO_SD_S;QN_6;MO_RR;MX_RS;eor
        282;18810101;18810131;    5;  -999;   -999;   -2.70;    -999;-999;-999;-999;-999;-999;-999;-999;-999;eor
        282;18810201;18810228;    5;  -999;   -999;    3.50;    -999;-999;-999;-999;-999;-999;-999;-999;-999;eor
        282;18810301;18810331;    5;  -999;   -999;    7.20;    -999;-999;-999;-999;-999;-999;-999;-999;-999;eor
        282;18810401;18810430;    5;  -999;   -999;   10.40;    -999;-999;-999;-999;-999;-999;-999;-999;-999;eor
        282;18810501;18810531;    5;  -999;   -999;   17.80;    -999;-999;-999;-999;-999;-999;-999;-999;-999;eor
```

(original data in .txt format)

The data is tabularly structured and its accuracy, completeness, consistency and timeliness is ensure by a trusted institution. The data is relevant as it contains temperature at 2 meters above ground level and the min and max temp for the month. Vineyards are heavily affected by ambient temperature, including extrema like freezing.

The license information can be found in German here and in English here. The weather data is offered under a Creative Commons 4.0 license. We will fulfill the obligations to name the data source used in a manner which complies with these guidelines.

Tolga Keskinoglu
Matriculation #: 23188690

# Data Pipeline

We used Python with the urllib.request library for downloading the dataset from the DWD website, the zipfile library to extract the .txt file from the downloaded ZIP file, and pandas for temporarily storing and transforming the data.

The following transformations were applied:

1. Loading the CSV (in .txt form) into pandas converted data types into appropriate values like int64 and float64
2. The "eor" (end of record) column was removed as that's not necessary for our purposes.
3. Missing or invalid data are marked with -999 in the dataset. We replace these with pd.NA (NA values) since we don't want to be able to compare, numerically, valid and invalid data at all.

We first attempted to use Jayvee to create the pipeline but ran into issues with formatting. Specifically, it was too difficult to remove (especially varying amounts of) preceding whitespace from a text file which separated values by with a character. This also prevented us from changing the value types from string to integer or decimal. We were also unable to chain multiple transformations, making this issue too time consuming to resolve with Jayvee.

The issue was resolved when we switched to Python using additional libraries.
The DWD has a fixed format for files and data, so we don't expect any deviation from the current format unless DWD standards change.

We do check to see if the file was downloaded, and the data extracted correctly. If not, the process is halted. It's worthy of note that the DWD also has a standardized file naming structure.

The data we are using is updated regularly (at least monthly). No assumptions were made regarding the size of the data we are processing so additional or updated rows should not affect the pipeline negatively. The addition of an entirely new category of data (a new column) would go unused but should not break the pipeline because of how pandas dataframes work.

Tolga Keskinoglu
Matriculation #: 23188690

# Result and Limitations

| | STATIONS_ID | MESS_DATUM_BEGINN | MESS_DATUM_ENDE | QN_4 | MO_N | MO_TT | MO_TX | MO_TN | MO_FK | MX_TX | MX_FX | MX_TN | MO_SD_S | QN_6 | MO_RR | MX_RS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1595 | 282 | 20131201 | 20131231 | 9 | <NA> | 3.03 | 5.97 | 0.29 | 1.87 | 11.8 | 19.6 | -5.8 | 46.0 | 9 | 24.2 | 3.4 |
| 994 | 282 | 19631101 | 19631130 | 5 | 6.25 | 7.21 | 10.98 | 3.42 | 2.01 | 18.3 | <NA> | -2.2 | 48.4 | 5 | 87.7 | 16.4 |
| 742 | 5705 | 19421101 | 19421130 | 5 | <NA> | <NA> | 6.8 | <NA> | <NA> | <NA> | <NA> | <NA> | <NA> | <NA> | <NA> | <NA> |
| 766 | 282 | 19441101 | 19441130 | <NA> | <NA> | <NA> | <NA> | <NA> | <NA> | <NA> | <NA> | <NA> | <NA> | 5 | 97.1 | <NA> |
| 952 | 5705 | 19610901 | 19610930 | 5 | 4.23 | 17.46 | 23.71 | 12.15 | 1.56 | 29.9 | <NA> | 5.9 | 192.7 | 5 | 34.9 | 11.6 |

(output data, displayed as a pandas dataframe)

The data is tabularly structured and its accuracy, completeness, consistency and timeliness is ensure by a trusted institution. The data is relevant as it contains temperature at 2 meters above ground level and the min and max temp for the month. Vineyards are heavily affected by ambient temperature, including extrema like freezing.

The output is formatted as a .sqlite file for long-term storage and wide compatibility. We intend to analyze the data using pandas dataframes.

We intend to compare these categories or features between the two datasets and see how much overlap there is.

We're conflicted on renaming the columns; on the one hand they are well-defined with precise meanings from DWD but on the other hand one needs to become familiar with this naming convention to work with them effectively.

We're also concerned that we'll run into errors when trying to compare periods with missing data. However, because a comparison is not possible with missing data points, we'll likely drop the rows (months) of the specific columns (features) which cannot be compared. We're unable to do this during the data pipeline because that would create a poorly structured database (some columns will be missing some months that other columns are not missing).