Name: Jyrki Aho
Student number:

AI part 2 assessment

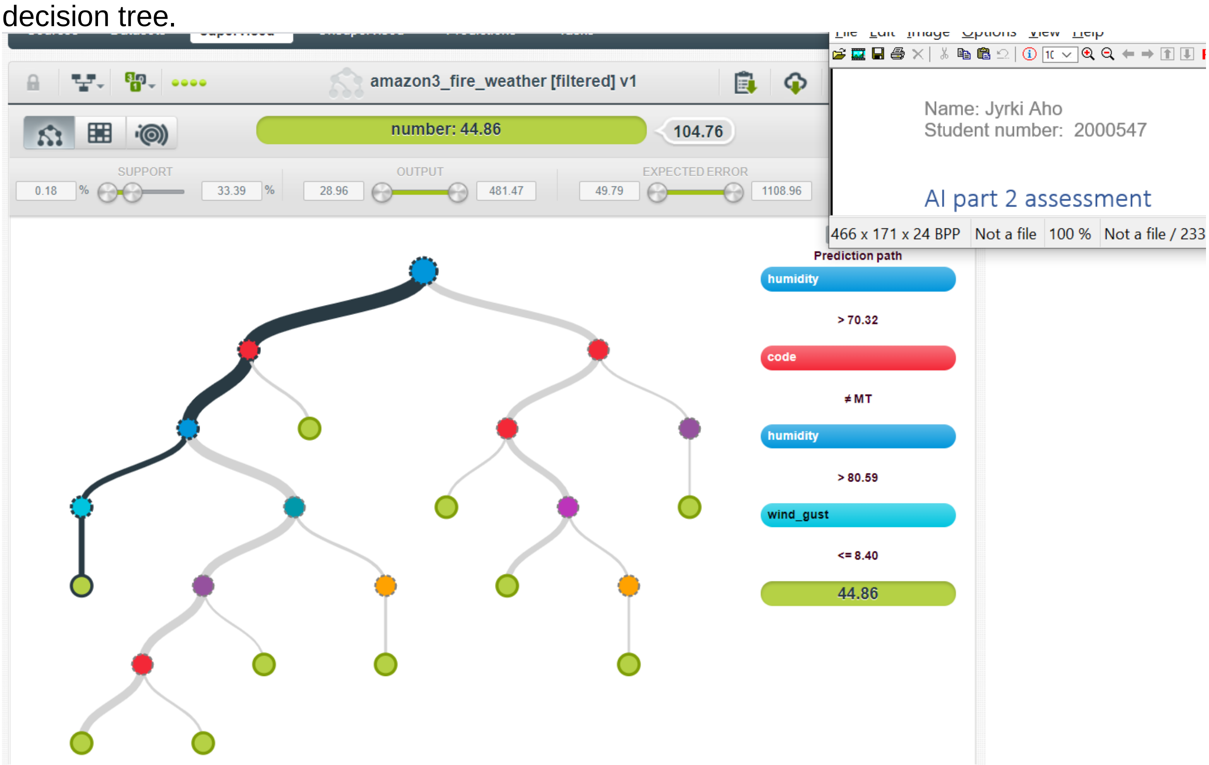
Brazilian forest fires did sound interesting data, but it did lack weather information. Because of that I did download Brazilian forest fire data from address https://www.kaggle.com/gustavomodelli/forest-fires-in-brazil. Because this data did not contain weather information, so I downloaded weather data from address https://www.kaggle.com/saraivaufc/automatic-weather-stations-brazil?select=automatic_weather_stations_inmet_brazil_2000_2021.csv. This did contain 612 weather stations hourly data from years 2000 to 2021.

Processing data

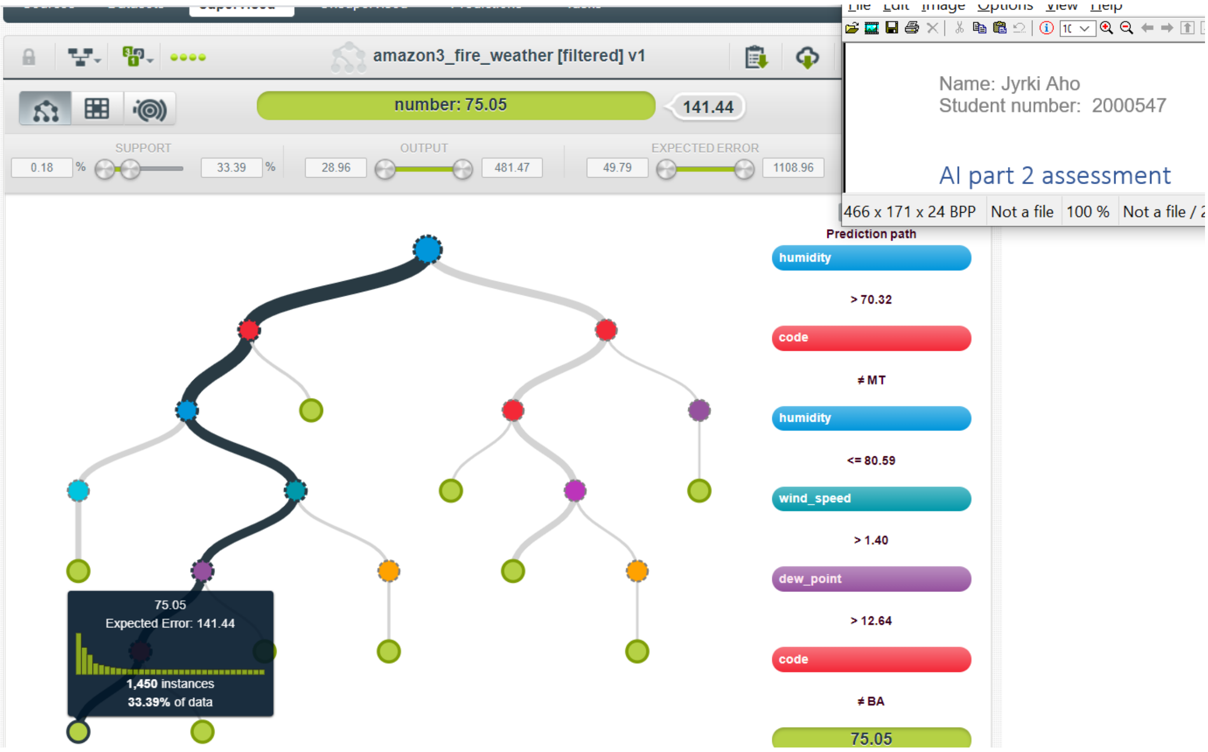
Based on the fores fire datasets dates, the data are gathered monthly and does not contain exact informations when the forest fire has started in which states. Because of this we have to calculate new weather data by grouping data based on month, year and states where the forest fire report has come. We have to also notice that forest fire data has been gathered from 1998 to 2007 and weather data from 2000 to 2021, we have to use inner join to combine these dataset to one big datasets. Because the weather dataset size are around 6 GB, so I did use Pythons Jupyter notebook to combine these datasets. I have added the Python code at the end of this document. I also have to use Python to combine these two datasets, because it was very slow process in Big ML, eventough I did use only one column to combine datasets.

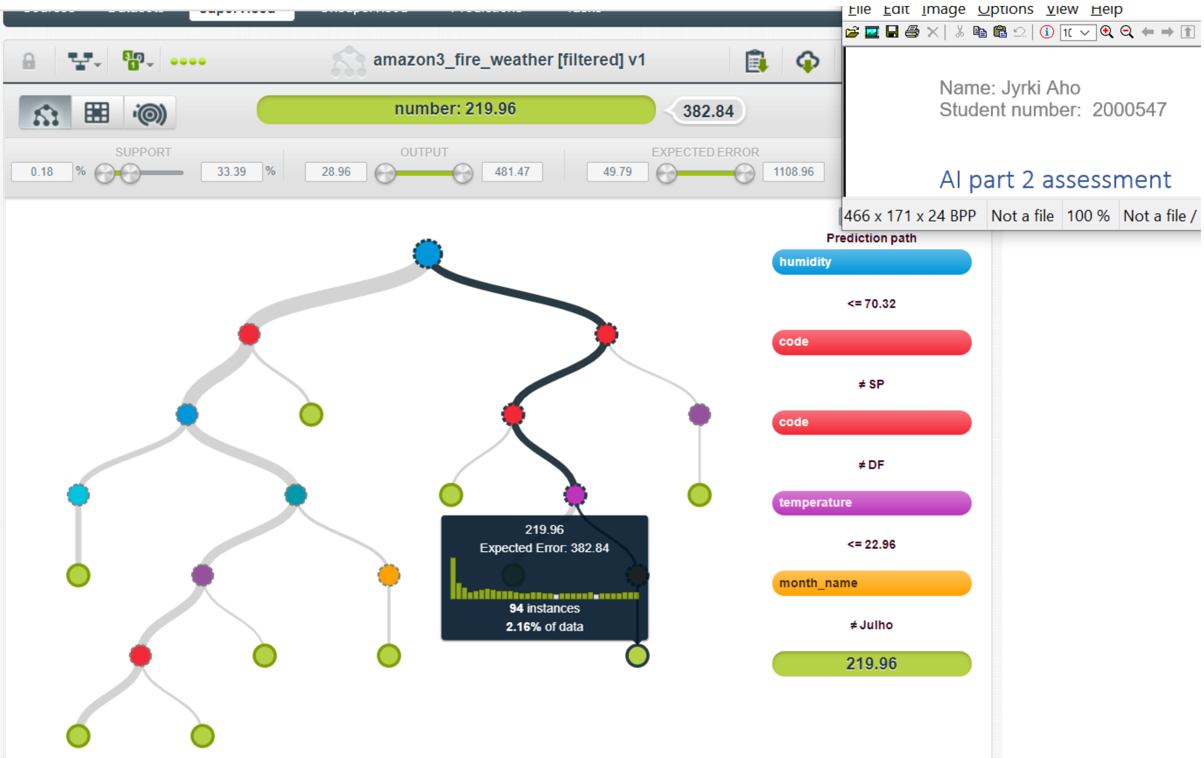
Decision tree

First I did try to model data with decision trees. Those decision trees did seems to have too many values to make predictions, so I did drop all the min and max values from the datasets. I also did take of month information. When I did choose Big ML to optimize data, the server did take some time to process data and to create tree. It did create following simplified decision tree.



Name: Jyrki Aho
Student number:





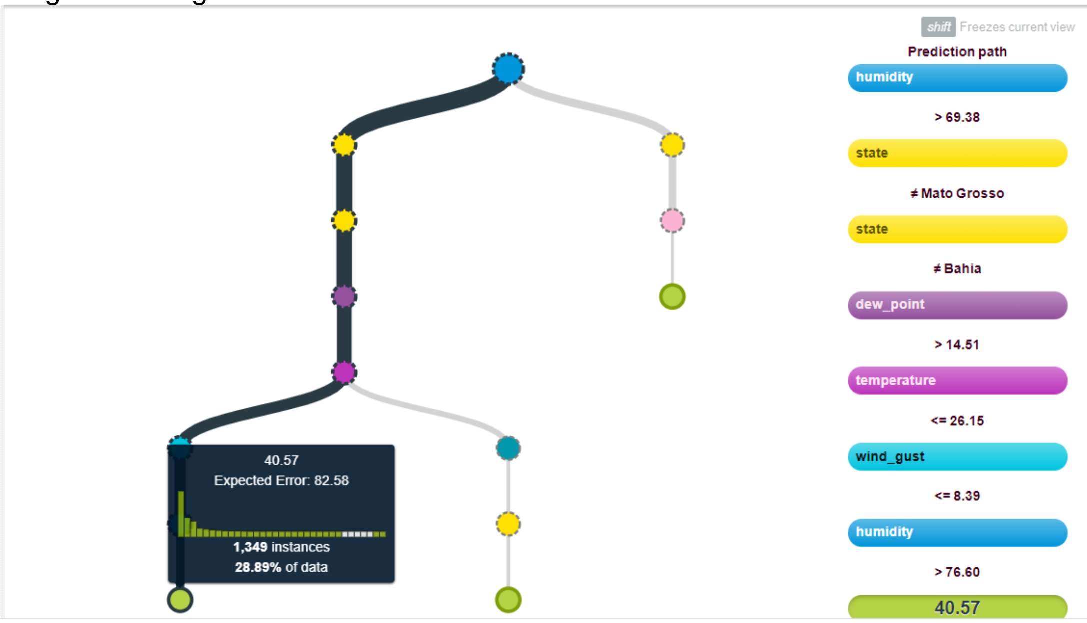
The model summary report did show column importance as
Field importance:

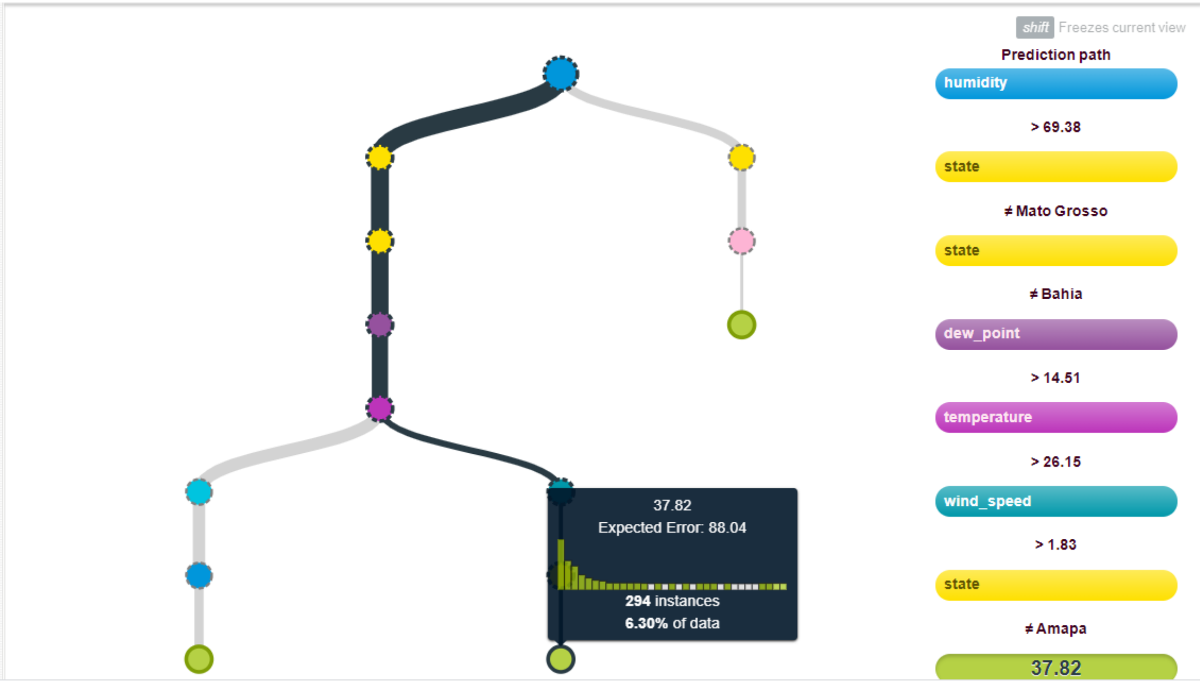
1. code: 40.42%
2. humidity: 24.25%
3. month_name: 9.75%
4. dew_point: 8.81%

Name: Jyrki Aho
Student number:

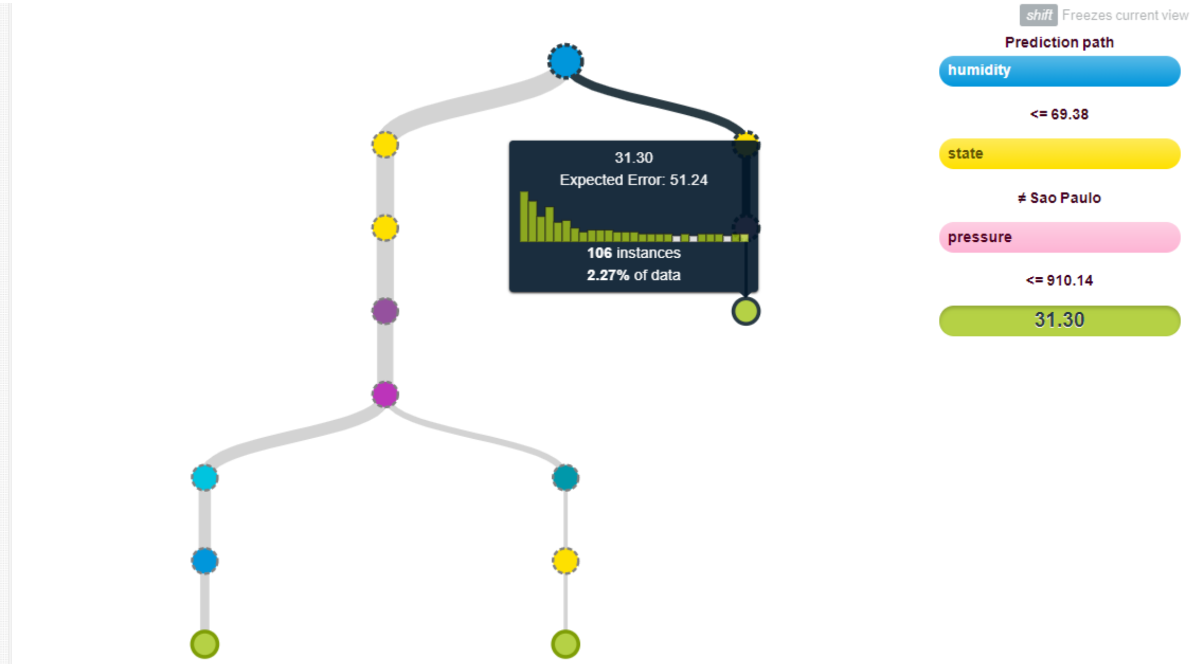
5. temperature: 6.98%
6. wind_speed: 6.03%
7. wind_gust: 3.76%

If we try other dataset where we drop min/max information, wind direction and code (which correlates with states). Then we can create new decision tree with optimizing data, and then we get following model.





Name: Jyrki Aho
Student number:



Then we get new field importance as
1. state: 35.08%
2. humidity: 16.58%
3. temperature: 12.73%
4. dew_point: 8.87%
5. month_name: 8.21%
6. pressure: 5.89%
7. wind_speed: 5.46%
8. wind_gust: 4.92%
9. radiation: 2.27%

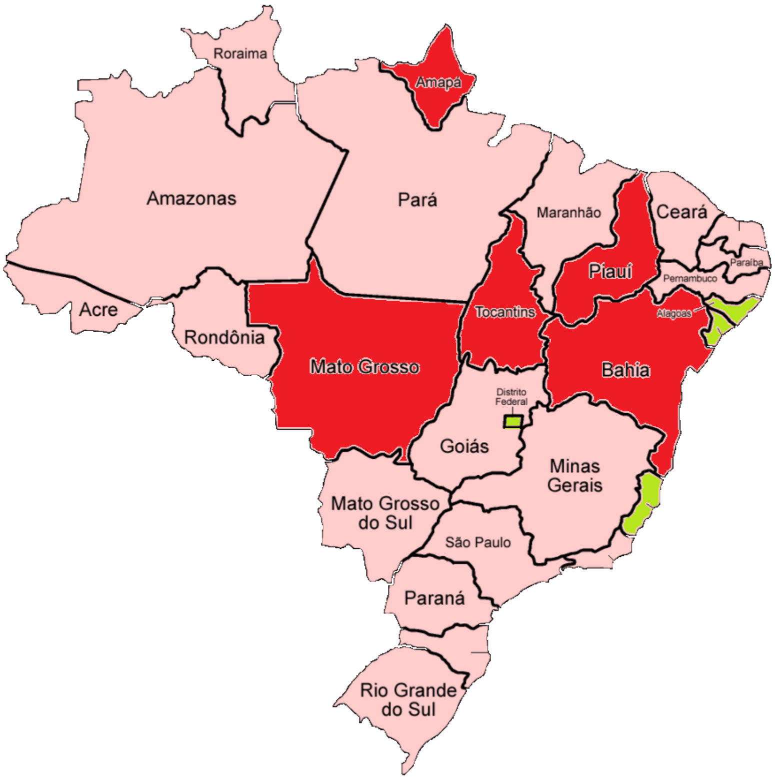
Using this information we could understand that some states have more forest fires than others. From this we could conclude that the vegetation and forest structure differs by state to state. This is understandable given the size of Brazil, in addition to information that the edges of rainforests are being burned in Brazil and new fields are being created in these places. It is also interesting information, that humidity does have effect of forest fire. According to the U.S. National Park Service, relative humidity is important because dead vegetation and air exchange moisture with each other. Low humidity absorbs water from dead vegetation and high humidity in turn transfers water to dead vegetation. In particular, small debris such as conifers react to changes faster than, for example dead branches. Months also have effect of forest fire, which probably means that some months are hotter and some a rainier, which probably has direct correlation to forest fires. It also seems that dew point, temperature and wind speed does have some effect to forest fires. Wind gust probably has some effect which helps fire to spread faster.

Linear regression

Linear regression model did not seem to work as I were expecting it to work. It does show information of the data, but I have slight problem to understand PDP graph. But when creating different dataset, we can see that some states have much more forest fires than others. Example Espirito Santo, Sergipe, Alagoas, Distrito Federal and Rio have much less forest fires than others. Instead Mato Grosso, Amapa, Tocantis, Piau and Bahia have much more forest fires. I did use Wikipedia pictures of Brazil (https://commons.wikimedia.org/wiki/File:Brazil_states_named.png), in which I did add green

Name: Jyrki Aho
Student number:

color to show states where are less forest fires and with red color states where are more forest fires than other states.



Linear model did also show that most likely forest fire occurs in July, October, November, and December. It also seems that higher temperature and precipitation lowers the risk of forest fires, but dew point increases the risk of forest fires. Dataset also did show that forest fires are slowly increasing.

Bias and predictors	Type	Coefficients
Bias	123	-1200.84000
state = Rio	ABC	-74.35980
state = Mato Grosso	ABC	117.56500
state = Paraiba	ABC	-11.68470
state = Amazonas	ABC	63.07200
state = Bahia	ABC	71.53580
state = Distrito Federal	ABC	-150.08500
state = Goias	ABC	28.16700
state = Sao Paulo	ABC	39.65750
state = Minas Gerais	ABC	12.31760
state = Alagoas	ABC	-91.73490
state = Maranhao	ABC	36.48640
state = Par�	ABC	32.02320
state = Santa Catarina	ABC	-44.45090

Name: Jyrki Aho
Student number:

Bias and predictors	Type	Coefficients
state = Par�	ABC	32.02320
state = Santa Catarina	ABC	-44.45090
state = Ceara	ABC	14.61980
state = Pernambuco	ABC	-27.91310
state = Piau	ABC	68.72790
state = Sergipe	ABC	-108.06400
state = Tocantins	ABC	62.01850
state = Espirito Santo	ABC	-108.52700
state = Rondonia	ABC	-6.12500
state = Acre	ABC	0
state = Amapa	ABC	51.54140
state = Roraima	ABC	-9.57977
month_name = Novembro	ABC	94.24430
month_name = Outubro	ABC	85.80970
month_name = Setembro	ABC	21.91230
month_name = Agosto	ABC	65.69370
month_name = Julho	ABC	63.12050
month_name = Junho	ABC	17.97980
month_name = Maio	ABC	-6.71640
month_name = Abril	ABC	0
month_name = Fevereiro	ABC	10.15760
month_name = Mar�o	ABC	15.82850
month_name = Janeiro	ABC	55.42970
month_name = Dezembro	ABC	62.33900
precipitation	123	-91.21960
pressure	123	0.20238
radiation	123	-0.00154
temperature	123	-59.86160
dew_point	123	49.48450
humidity	123	-15.64680
wind_gust	123	-14.72850
wind_speed	123	28.37980
year	123	1.39488
precipitation	missing	-27.43590
pressure	missing	37.19830
radiation	missing	35.86970
temperature	missing	-1549.24000
dew_point	missing	0

Name: Jyrki Aho
Student number:

Sources

US National Park Service. 2021. Understanding fire danger. Can be read at: https://www.nps.gov/articles/understanding-fire-danger.htm. Read: 16.7.2021.

Python code

```
#Handling forest fire data
import pandas as pd
import numpy as np
import matplotlib as plt

amazon_df = pd.read_csv("amazon.csv")
print(len(amazon_df))                                    # 6 454 datapoints


# change month to numbers
month_map={'Janeiro': 1, 'Fevereiro': 2, 'Mar�o': 3, 'Abril': 4, 'Maio': 5,
        'Junho': 6, 'Julho': 7, 'Agosto': 8, 'Setembro': 9, 'Outubro': 10,
        'Novembro': 11, 'Dezembro': 12}
amazon_df['month_nbr']=amazon_df['month'].map(month_map)
amazon_df['yearmonth'] = amazon_df['year']*100 + amazon_df['month_nbr']
amazon_df.head(50)

# Check for NaN values
print( amazon_df['month_nbr'].isnull().values.any() ) # False

k = amazon_df['state'].unique()
k.sort()
print(k) # contains name of 23 Brazilian states


# creating are codelist
sc_df = pd.read_csv("automatic_stations_codes_2000_2021.csv",delimiter=";")
sc_df.head(20)

codigo  = sc_df['CODIGO']
uf   = sc_df['UF']
state_map2 = dict(zip(codigo,uf))
print(state_map2)
print(len(codigo))


#creating state map list
amazon_df['code']=amazon_df['state'].map(state_map)
print(amazon_df)


#Save data to file
amazon_df.to_csv("amazon2.csv",sep=';')
```

Name: Jyrki Aho
Student number:

```
#Handling weather data using previous lists
weather_df = pd.read_csv("automatic_weather_stations_inmet_brazil_2000_2021.csv",
delimiter=";")
weather_df.head(40)

n = len(weather_df)
print(n)                    # contains 60 452 376 data points

#Contains multiple difficult column names, which have to simplify
h = weather_df.columns
#print(h)
new_header=['station', 'date', 'hour', 'precipitation', 'pressure', 'max_preasure',
'min_preasure',
            'radiation', 'temperature', 'dew_point', 'max_temperature', 'min_temperature',
           'max_dew_point', 'min_dew_point', 'max_humidity', 'min_humidity', 'humidity',
'wind_direction',
           'wind_gust', 'wind_speed']
z = dict(zip(h,new_header))
weather_df= weather_df.rename(columns=z)
#weather_df.head(20)


# data contains NaN values, which are ok, but -9999 values have to change to NaN values
cols = ['precipitation', 'pressure', 'max_preasure', 'min_preasure',
            'radiation', 'temperature', 'dew_point', 'max_temperature', 'min_temperature',
           'max_dew_point', 'min_dew_point', 'max_humidity', 'min_humidity', 'humidity',
           'wind_gust', 'wind_speed']
for c in cols:
    weather_df[c]= weather_df[c].apply(lambda x: np.nan if x<-1000 else x)


#lets create yearmonth column
weather_df['year'] = [int(x[0:4]) for x in weather_df['date']]
weather_df['month'] = [int(x[5:7]) for x in weather_df['date']]
weather_df['yearmonth'] = [(100*int(x[0:4]) + int(x[5:7])) for x in weather_df['date']]


#Lets add area code column to data
weather_df['code']=weather_df['station'].map(state_map2)
weather_df.head(20)


# Now we will create new dataframes and then we combine weather data to one file
dataset_mean = weather_df.groupby(['yearmonth', 'code']).mean()
dataset_max =  weather_df.groupby(['yearmonth','code']).max()
dataset_min =  weather_df.groupby(['yearmonth','code']).min()


dataset_weather = dataset_mean.copy()
dataset_weather.drop(columns=['hour'])


dataset_weather['max_preasure'] = dataset_max['max_preasure']
dataset_weather['min_preasure'] = dataset_min['min_preasure']
```

Name: Jyrki Aho
Student number:

```
dataset_weather['max_temperature'] = dataset_max['max_temperature']
dataset_weather['min_temperature'] = dataset_min['min_temperature']

dataset_weather['max_dew_point'] = dataset_max['max_dew_point']
dataset_weather['min_dew_point'] = dataset_min['min_dew_point']

dataset_weather.head(20)

dataset_weather.to_csv("amazon_weather.csv", sep=";")



#Lets now combine brazilian fire and weather data

amazon_weather = pd.read_csv("amazon_weather.csv", delimiter=";")
amazon_weather.drop(amazon_weather.columns[amazon_weather.columns.str.contains('un
named',case = False)],axis = 1, inplace = True)
amazon_weather.head(10)


a = amazon_weather['code']
b = amazon_weather['yearmonth']
amazon_weather['id'] =amazon_weather[['code','yearmonth']].astype(str).apply(''.join,1)
amazon_weather= amazon_weather.drop(columns=['hour','yearmonth'])
amazon_weather.head(10)


amazon_fire = pd.read_csv("amazon2.csv", delimiter=";")
amazon_fire.drop(amazon_fire.columns[amazon_fire.columns.str.contains('unnamed',case =
False)],axis = 1, inplace = True)
amazon_fire.head(10)



amazon_fire['id'] =amazon_fire[['code','yearmonth']].astype(str).apply(''.join,1)
amazon_fire = amazon_fire.drop(columns=['year', 'month_nbr','yearmonth','date','code'])
amazon_fire.head(10)



amazon3 = pd.merge(amazon_fire, amazon_weather,  how='inner', left_on='id', right_on
='id')
amazon3.head(20)


amazon3.to_csv("amazon3_fire_weather.csv",sep=";",index=False)
```