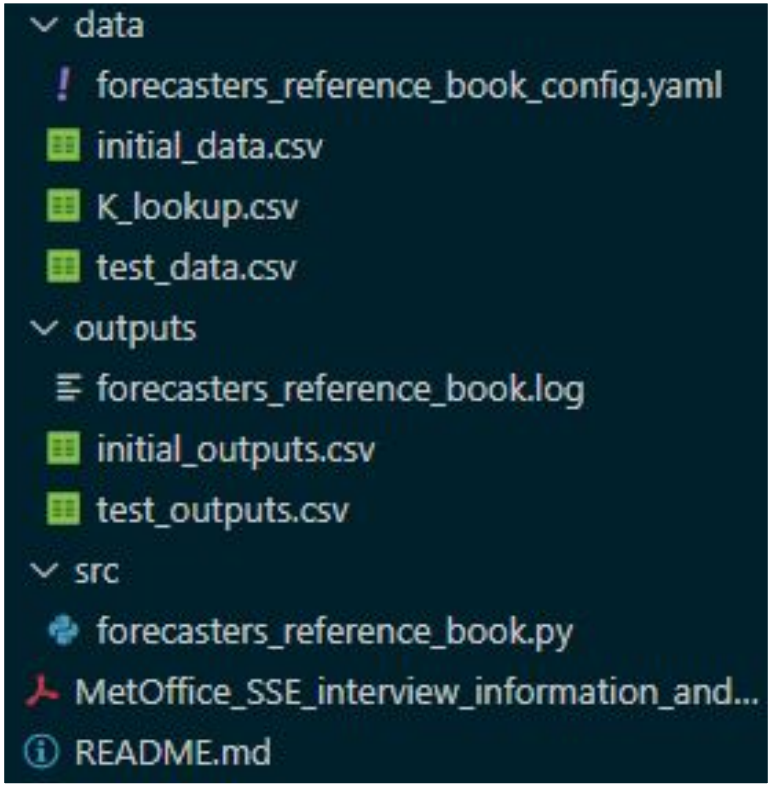
Scientific Software Engineer Exercise

Joshua Minien

Solution outline

1. Data Science project structure
2. Data files
3. Forecaster's reference book programme built in Python3
4. Further ideas

Data science project structure



Data files

1. Configuration
2. K value lookup data sheet
3. Data to be processed for forecasting

Configuration

```yaml
method_data:
  k_lookup_file_path: "../data/K_lookup.csv"

data:
  data_file_path: "../data/initial_data.csv"

outputs:
  decimal_place_precision: 2
  output_filename: "initial_outputs.csv"
  output_directory: "../outputs/"
```

K value lookup data sheet

wind speed min. (knots)	wind speed max. (knots)	cloud cover min. (oktas)	cloud cover max. (oktas)	K ()
0	12	0	2	-2.2
0	12	2	4	-1.7
0	12	4	6	-0.6
0	12	6	8	0
13	25	0	2	-1.1
13	25	2	4	0
13	25	4	6	0.6
13	25	6	8	1.1
26	38	0	2	-0.6
26	38	2	4	0
26	38	4	6	0.6
26	38	6	8	1.1
39	51	0	2	1.1
39	51	2	4	1.7
39	51	4	6	2.8
39	51	6	8	NaN

Data to be processed for forecasting

```
Temp. noon (celcius),Temp. dew point noon (celius),wind speed (knots),cloud cover (oktas),location,date
22.4,10.9,14.56,3.9,A,1
18.6,12.56,3.4,6,B,1
26,8.5,0,0.0,B,2
13.2,9.4,12.5, 4.1,C,2
```

Forecaster's reference book programme

1. For simplicity I have only added INFO level logging
2. Imported modules: numpy, pandas, yaml
3. YAML file imports for configuration
4. K value imported from .csv file as a pandas DataFrame with correct data types
5. Process data imported from .csv file as a pandas DataFrame with correct data types
6. Uses numpy arrays to efficiently find the relevant K values for data
7. Lambda function on imported data DataFrame to evaluate the minimum temperature at 12 pm
8. Write appended DataFrame to an output .csv file

Code demo

Forecaster's reference book programme- Test output

Date	Location	Midday Temperature (℃)	Midday Dew Point (℃)	Wind (Kn)	Cloud (oktas)	Forecasted Minimum Temperature (℃)
1	A	22.4	10.9	14.56	3.9	11.81
1	B	18.6	12.65	3.4	6	10.97
2	B	26	8.5	0	0.0	9.43
2	C	13.2	9.4	12.5	4.1	6.38

Further ideas

1. Add unit tests for programme such that it could be used in a CI/CD pipeline
2. If in a large code base:
    a. Write function for the computation that uses numpy arrays
    b. Write function to a module
    c. Import module where relevant in the code base
3. Implement more robust logging and error handling for team coding and use a VCS
4. Code uses vectorisation ready for larger data inputs
5. We do not have raw data to compare; if available import these data to conduct a statistical test for testing accuracy of historical method
6. Development team for the work, assuming automations are in place:
    a. Scientist requesting code
    b. Developer
    c. Code reviewer