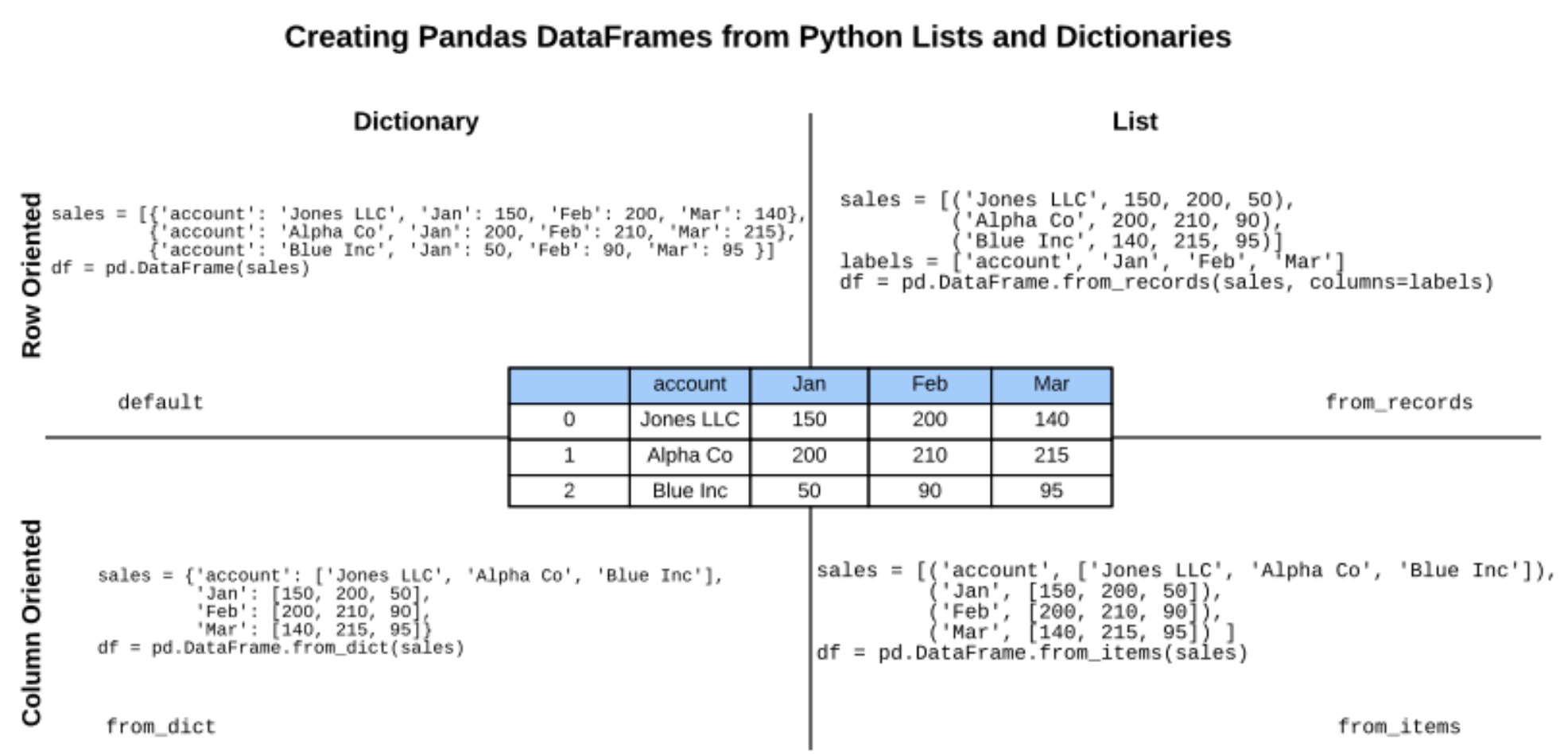
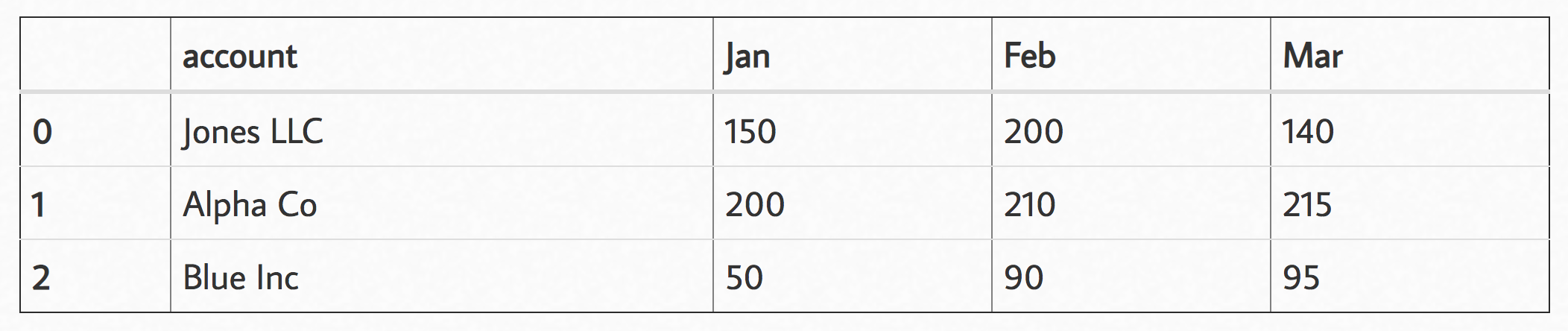
Hands-On Practices for Python Data Analysis Basics

1. 上網搜尋哭臉的萬國碼(16進位)，以Python查詢其code point的10進位值，並轉換為utf8的位元組值。(搜尋關鍵字：哭臉 unicode)
2. Python提供了許多有用的資料結構，說明下面資料結構何者為原生型？何者為衍生型？及其間的差別：Tuples, Lists, Dictionaries, Set, Arrays, Series, DataFrames，並與R語言中的資料結構Vector, Matrix, Array, List, DataFrames對比，找出最接近的R資料結構。(Python vs. R: Lists vs. Vector, Matrix, Array; Dictionaries: List; Arrays vs. Vector, Matrix, Array; Series vs. Vector; DataFrames vs. DataFrames)
3. 練習載入定位符分隔的檔案data1.tab，載入過程與逗號分隔檔案.csv完全相同。
4. 分別創建兩個串列內含三個整數(假設分別為1, 2, 3和4, 5, 6)，將其用二元的加號運算子進行運算。
5. 載入numpy模組並簡記為np，將上題的兩個串列轉成numpy的陣列物件，再用二元的加號運算子進行運算，並將結果與上題比較。
6. 延續上題，運用dot方法將兩numpy陣列物件進行點積運算。
7. 載入TRD\_Index.txt，對資料集中上海綜合證券指數收益率與深圳綜合證券指數收益率進行相關性分析。
8. 參照上面兩表，練習以橫列或縱行為導向、以及運用字典或是串列的原生資料結構，建立中間的pandas DataFrame表，請思考有無領悟到其間的差異。
9. 載入pandas模組並簡記為pd，創建下列男女性觀看電視時數的DataFrame，練習連續位置取值與間斷位置取值的pandas資料框語法，並進行男女群組與摘要的計算分析。

Gender TV

0 f 3.4

1 f 3.5

2 m 2.6

3 f 4.7

4 m 4.1

5 m 4.1

6 f 5.1

7 m 3.9

8 f 3.7

9 m 2.1

10 m 4.3

10. 請問nba\_2013.csv是寬資料還是長資料？根據上面的答案將之轉為另一種格式。

11. 請將nba\_2013.csv的資料，依球星得分(FG)從高到低排序整個資料表。

12. 載入R語言datasets套件中的紐約空氣品質資料集airquality，並檢視其整體與行列的遺缺狀況。

13. 請以Kolmogorov-Smirnov檢定x是否來自常態分佈。

import scipy.stats as stats

import random

import numpy as np

x = np.random.randn(100)

14. 請運用Python統計函數，計算12 ~ 20的眾數。

15. 請接續以下程式，完成Simple linear regression 的分析

import pandas as pd

df1 = pd.read\_stata('https://github.com/QuantEcon/QuantEcon.lectures.code/raw/master/ols/maketable1.dta')

df1.head()

16. 請將IQ資料集(examples/brain\_size.csv)分成男和女，進行t-test和成對t-test分析。原始資料請至<http://www.scipy-lectures.org/_downloads/brain_size.csv> 下載。

17. 請根據澳洲墨爾本1981-1990十年間每日最低溫度的資料，建立AR 自我迴歸模型(Autoregressive Models)。原始資料集請至[https://datamarket.com/data/set/2324/daily-minimum-temperatures-in-melbourne-australia-1981-1990#!ds=2324&display=line](https://datamarket.com/data/set/2324/daily-minimum-temperatures-in-melbourne-australia-1981-1990%25252525252523!ds=2324&display=line) 下載。

18. sklearn.datasets中有一鳶尾花iris資料集，請進行主成份降維分析，並說明妳/你所得到的結果。  
19. (整合題)請將電信業常用的客戶流失建模資料集churn.csv載入Python環境中，練習

A. 資料理解與遺缺值辨識

B. 挑出訓練集(training set)；

C. 建立屬性矩陣(feature matrix)；

D. 區分類別與數值屬性；

E. 建立類別標籤向量(class label)；

F. 低變異過濾(low variance filter)；

G. 偏斜(skewed)分佈屬性Box-Cox轉換；

H. 主成份分析維度縮減(dimensionality reduction)；

I. 高相關過濾。

J. K-Means集群