ML Final Project

Rainfall Forecast

Team 18

- 0716032 林佑鑫
- 0716050 吳泓毅
- 0716070 張育維

Motivation

- Rains a lot recently
- Often forget bringing umbrella

Goals

- Remind us to bring umbrellas if it is going to rain.

Data Collection

- Use Web Crawler to fetch public data from the CWB website
- Parse the data from different webpage and join them together

月報表 (monthly data)　測站:466920_臺北　466920_臺北　觀測時間:2020-11　CSV下載　資料定義請詳見 ◎網頁說明Readme

	press						temperature						Dew Point	RH			WS/WD						Prec			
觀測時間 (day)	測站氣壓 (hPa)	海平面氣壓 (hPa)	測站最高氣壓 (hPa)	測站最高氣壓時間 (LST)	測站最低氣壓 (hPa)	測站最低氣壓時間 (LST)	氣溫 (℃)	最高氣溫 (℃)	最高氣溫時間 (LST)	最低氣溫 (℃)	最低氣溫時間 (LST)	露點溫度 (℃)	相對濕度 (%)	最小相對濕度 (%)	最小相對濕度時間 (LST)	風速 (m/s)	風向 (360degree)	最大陣風 (m/s)	最大陣風風向 (360degree)	最大陣風風速時間 (LST)	降水量 (mm)	降水時數 (hour)	最大十分鐘降水量 (mm)	最大十分鐘降水量起始
ObsTime	StnPres	SeaPres	StnPresMax	StnPresMaxTime	StnPresMin	StnPresMinTime	Temperature	T Max	T Max Time	T Min	T Min Time	Td dew point	RH	RHMin	RHMinTime	WS	WD	WSGust	WDGust	WGustTime	Precp	PrecpHour	PrecpMax10	PrecpMax10
01	1011.3	1014.9	1013.3	2020-11-01 09:15	1009.3	2020-11-01 14:28	25.2	29.8	2020-11-01 12:56	22.3	2020-11-01 03:53	19.0	70	44	2020-11-01 13:09	3.5	80	13.2	70	2020-11-01 11:36	T	0.8	T	2020-11-01
02	1011.9	1015.4	1014.4	2020-11-02 21:07	1010.2	2020-11-02 04:06	24.6	28.3	2020-11-02 10:56	22.3	2020-11-02 21:21	18.8	71	50	2020-11-02 14:41	3.4	80	14.8	40	2020-11-02 14:45	0.0	0.0	0.0	
03	1015.5	1019.1	1017.4	2020-11-03 20:36	1013.7	2020-11-03 00:31	21.6	24.9	2020-11-03 11:58	18.7	2020-11-03 23:49	16.0	72	45	2020-11-03 12:06	3.0	80	11.2	70	2020-11-03 12:22	0.5	0.1	0.5	2020-11-03
04	1016.0	1019.6	1017.8	2020-11-04 09:05	1014.1	2020-11-04 14:11	21.8	25.0	2020-11-04 13:13	18.7	2020-11-04 00:01	14.3	63	42	2020-11-04 13:23	3.5	80	13.8	70	2020-11-04 11:59	0.0	0.0	0.0	
05	1014.3	1017.9	1016.4	2020-11-05 09:25	1012.4	2020-11-05 15:02	22.9	26.0	2020-11-05 10:54	21.4	2020-11-05 01:16	17.9	74	54	2020-11-05 10:55	3.6	80	12.6	60	2020-11-05 15:22	0.0	0.0	0.0	
06	1009.4	1012.9	1013.3	2020-11-06 00:01	1006.3	2020-11-06 15:26	27.0	31.3	2020-11-06 14:33	22.5	2020-11-06 00:01	22.8	79	54	2020-11-06 14:35	4.2	80	15.9	130	2020-11-06 23:09	0.5	0.3	0.5	2020-11-06
07	1010.6	1014.1	1014.7	2020-11-07 23:36	1007.0	2020-11-07 01:12	27.5	32.6	2020-11-07 10:52	22.8	2020-11-07 22:59	21.2	70	43	2020-11-07 11:07	2.9	80	16.2	130	2020-11-07 00:37	0.0	0.0	0.0	
08	1016.8	1020.4	1018.5	2020-11-08 20:46	1014.5	2020-11-08 00:01	20.9	22.9	2020-11-08 00:02	19.7	2020-11-08 21:53	17.8	83	71	2020-11-08 11:50	2.9	90	11.7	80	2020-11-08 11:53	0.0	0.0	0.0	
09	1017.4	1021.0	1018.6	2020-11-09 20:29	1016.0	2020-11-09 14:50	20.8	21.7	2020-11-09 10:58	19.6	2020-11-09 03:36	15.2	70	62	2020-11-09 14:51	3.5	90	12.1	80	2020-11-09 10:43	0.0	0.0	0.0	
10	1018.5	1022.1	1020.0	2020-11-10 20:21	1016.7	2020-11-10 03:23	20.5	21.7	2020-11-10 11:59	19.0	2020-11-10 05:06	15.8	75	62	2020-11-10 23:15	3.1	90	13.8	60	2020-11-10 11:46	0.5	2.3	0.5	2020-11-10
11	1019.5	1023.1	1021.0	2020-11-11 09:08	1018.4	2020-11-11 03:24	22.5	25.6	2020-11-11 11:09	20.6	2020-11-11 01:52	15.0	63	46	2020-11-11 10:57	4.2	80	15.1	70	2020-11-11 11:44	T	0.4	T	2020-11-11
12	1016.9	1020.5	1018.9	2020-11-12 00:01	1015.4	2020-11-12 15:44	22.0	24.4	2020-11-12 11:23	20.9	2020-11-12 03:15	18.8	82	61	2020-11-12 11:26	4.1	80	14.1	70	2020-11-12 12:12	0.5	2.3	0.5	2020-11-12
13	1015.7	1019.3	1017.3	2020-11-13 22:26	1014.0	2020-11-13 15:41	21.9	23.4	2020-11-13 15:12	20.9	2020-11-13 05:37	21.2	96	84	2020-11-13 12:52	3.0	80	10.7	70	2020-11-13 03:18	11.0	10.2	3.5	2020-11-13
14	1016.8	1020.4	1018.1	2020-11-14 09:36	1015.1	2020-11-14 03:29	21.9	23.4	2020-11-14 10:43	20.6	2020-11-14 06:43	18.7	83	69	2020-11-14 17:26	3.7	80	13.2	90	2020-11-14 13:56	4.0	0.6	1.0	2020-11-14
15	1015.5	1019.1	1017.0	2020-11-15 00:04	1013.9	2020-11-15 14:14	24.8	28.8	2020-11-15 13:50	21.8	2020-11-15 02:13	19.0	71	48	2020-11-15 15:20	3.7	80	11.5	100	2020-11-15 12:13	0.0	0.0	0.0	
16	1013.8	1017.3	1015.8	2020-11-16 09:33	1011.9	2020-11-16 15:08	25.2	29.0	2020-11-16 14:56	22.5	2020-11-16 05:04	19.7	73	53	2020-11-16 10:45	3.6	80	12.1	80	2020-11-16 15:42	0.0	0.0	0.0	
17	1011.8	1015.3	1013.5	2020-11-17 09:03	1009.8	2020-11-17 23:59	25.8	29.4	2020-11-17 11:57	24.1	2020-11-17 04:18	19.5	69	49	2020-11-17 11:42	4.1	80	11.7	60	2020-11-17 11:30	0.0	0.0	0.0	
18	1009.2	1012.7	1011.2	2020-11-18 09:13	1007.3	2020-11-18 14:34	26.7	32.8	2020-11-18 13:23	22.3	2020-11-18 06:24	19.9	68	39	2020-11-18 11:41	1.1	110	7.4	100	2020-11-18 00:40	0.0	0.0	0.0	
19	1008.5	1012.0	1010.7	2020-11-19 09:14	1006.4	2020-11-19 14:44	26.9	32.7	2020-11-19 11:48	22.9	2020-11-19 05:17	19.8	67	38	2020-11-19 10:54	1.5	190	6.8	70	2020-11-19 15:23	0.0	0.0	0.0	
20	1010.4	1013.9	1012.2	2020-11-20 20:09	1008.1	2020-11-20 00:51	25.4	30.4	2020-11-20 10:22	23.7	2020-11-20 05:21	20.9	77	48	2020-11-20 10:24	2.4	80	11.7	90	2020-11-20 16:34	0.0	0.0	0.0	
21	1012.3	1015.9	1013.9	2020-11-21 08:56	1011.1	2020-11-21 02:37	24.0	25.3	2020-11-21 12:51	22.7	2020-11-21 06:39	22.1	89	78	2020-11-21 14:36	3.5	80	10.4	80	2020-11-21 00:41	0.5	3.1	0.5	2020-11-21
22	1014.7	1018.2	1018.8	2020-11-22 21:31	1011.1	2020-11-22 00:06	24.6	30.3	2020-11-22 10:53	21.4	2020-11-22 23:44	20.4	79	47	2020-11-22 10:41	1.9	80	12.0	100	2020-11-22 20:53	0.0	0.0	0.0	
23	1017.8	1021.4	1019.7	2020-11-23 08:55	1016.0	2020-11-23 14:12	21.8	23.4	2020-11-23 11:40	20.8	2020-11-23 04:38	19.1	85	73	2020-11-23 09:42	4.3	90	14.5	90	2020-11-23 15:58	0.0	0.0	0.0	
24	1017.0	1020.5	1018.8	2020-11-24 09:20	1015.1	2020-11-24 14:39	22.9	25.9	2020-11-24 11:19	21.4	2020-11-24 02:21	18.4	76	56	2020-11-24 11:22	4.3	80	14.6	100	2020-11-24 12:15	0.0	0.0	0.0	
25	1015.2	1018.7	1016.8	2020-11-25 00:01	1013.1	2020-11-25 14:47	24.6	28.6	2020-11-25 10:56	22.2	2020-11-25 05:52	19.2	73	48	2020-11-25 10:34	4.4	80	12.7	120	2020-11-25 13:22	0.0	0.0	0.0	

```python
import csv
import time
import requests
from bs4 import BeautifulSoup
from datetime import date, timedelta

def daterange(start_date, end_date):
    for n in range(int((end_date - start_date).days)):
        yield start_date + timedelta(n)

start_date = date(2010, 1, 1) #from 2010-01-01
end_date = date.today() #today
csvfile="data.csv"

with open(csvfile,"w+",newline="",encoding="utf-8") as fp:
    writer=csv.writer(fp)
    for single_date in daterange(start_date, end_date):
        if single_date.strftime("%d") == '01':
            start_time = time.time()
            url="https://e-service.cwb.gov.tw/HistoryDataQuery/MonthDataController.do?command=viewMain&station=466920&stname=%25E8%2587%25BA%25E5%258C%2597&da
            r=requests.get(url)
            r.encoding="utf-8"
            soup=BeautifulSoup(r.text,"lxml")
            tag_table=soup.find(id="MyTable")
            rows=tag_table.findAll("tr")
            if single_date == start_date:
                for rnum, row in enumerate(rows):
                    rowList = []
                    for cnum, cell in enumerate(row.findAll(["td","th"])):
                        if rnum >= 3 and cnum == 0:
                            rowList.append(single_date.strftime("%Y-%m-")+cell.get_text().replace("\n","").replace("\r","").replace(u'\xa0', u'').replace(u'
                        elif rnum >= 2:
                            rowList.append(cell.get_text().replace("\n","").replace("\r","").replace(u'\xa0', u'').replace(u' ', u''))
                    writer.writerow(rowList)
            else:
                for row in rows[3:]:
                    rowList = []
                    for cnum, cell in enumerate(row.findAll(["td","th"])):
                        if cnum == 0:
                            rowList.append(single_date.strftime("%Y-%m-")+cell.get_text().replace("\n","").replace("\r","").replace(u'\xa0', u'').replace(u'
                        else:
                            rowList.append(cell.get_text().replace("\n","").replace("\r","").replace(u'\xa0', u'').replace(u' ', u''))
                    writer.writerow(rowList)
            print(single_date.strftime("%Y-%m") + ' finish!  cost time: %.2f seconds' %float(time.time() - start_time))
```

Data Content

- The daily weather data of Taipei station in Taipei.
- Features: Temperature, Sunshine, UV intensity, Cloud Amount, Wind, etc.
- Target: Precipitation Hours

ObsTime	StnPres	SeaPres	StnPresMax	StnPresMaxTime	StnPresMin	StnPresMinTime	Temperature	TMax	TMaxTime	TMin	TMinTime	Tddewpoint	RH	RHMin	RHMinTime	WS	WD	WSGust	WDGust	WGustTime	Precp	Precp
2010/1/1	1018.2	1019	1021.7	2010-01-0100:01	1015.7	2010-01-0114:24	16.5	19.4	2010-01-0113:16	13.7	2010-01-0100:01	11.1	71	65	2010-01-0110:18	3.6	70	13.4	70	2010-01-0111:30	0	
2010/1/2	1016.4	1017.2	1018.9	2010-01-0223:13	1013.8	2010-01-0214:09	16.8	19.6	2010-01-0212:10	14.9	2010-01-0223:40	14.1	84	74	2010-01-0211:53	1.1	160	8.4	330	2010-01-0215:27	4.5	
2010/1/3	1016.8	1017.6	1019.3	2010-01-0308:52	1015.3	2010-01-0323:55	15.8	18.2	2010-01-0312:22	14.1	2010-01-0306:31	13.3	85	76	2010-01-0318:50	2.3	60	9.8	60	2010-01-0320:40	9	
2010/1/4	1015.3	1016.1	1017	2010-01-0421:23	1012.9	2010-01-0414:19	17.5	22.2	2010-01-0414:23	13.8	2010-01-0405:53	13.7	79	66	2010-01-0414:53	1.5	70	7.4	60	2010-01-0416:55	0	
2010/1/5	1018.9	1019.7	1021.6	2010-01-0509:48	1016.7	2010-01-0500:32	14.9	17.8	2010-01-0501:18	13.5	2010-01-0520:07	11.9	82	76	2010-01-0503:43	4.4	70	15.4	70	2010-01-0511:40	0.2	
2010/1/6	1020.2	1021	1022.3	2010-01-0620:06	1017.6	2010-01-0603:01	14.4	15.1	2010-01-0614:11	13.6	2010-01-0623:23	12.6	89	84	2010-01-0616:15	3.5	80	10	70	2010-01-0621:57	5.5	
2010/1/7	1021.9	1022.7	1024.4	2010-01-0709:32	1020.4	2010-01-0715:05	13.5	13.9	2010-01-0721:11	13.1	2010-01-0706:27	12.2	91	88	2010-01-0700:11	3.6	70	10.5	50	2010-01-0706:17	19	
2010/1/8	1020.4	1021.2	1022.2	2010-01-0810:26	1018.8	2010-01-0803:50	13.6	14.7	2010-01-0823:29	13	2010-01-0804:35	12.4	92	86	2010-01-0823:29	2.4	90	8.7	100	2010-01-0800:39	24	
2010/1/9	1018.2	1019	1020.1	2010-01-0900:02	1015.6	2010-01-0914:56	18.2	23.3	2010-01-0914:09	14.4	2010-01-0901:47	14.3	79	60	2010-01-0914:56	2.7	90	9.7	90	2010-01-0900:22	T	
2010/1/10	1017.5	1018.3	1019.7	2010-01-1007:17	1014.4	2010-01-1012:55	18.5	25.5	2010-01-1013:11	15.4	2010-01-1007:00	15	81	51	2010-01-1013:02	1.7	290	10.7	310	2010-01-1013:21	1.9	
2010/1/11	1017.7	1018.5	1020.3	2010-01-1123:32	1015.8	2010-01-1114:50	16.5	19.3	2010-01-1112:28	13	2010-01-1123:49	13.1	81	72	2010-01-1115:44	1.7	30	9.2	40	2010-01-1123:25	4.4	
2010/1/12	1023.3	1024.1	1027	2010-01-1223:37	1020.3	2010-01-1200:01	10.5	13.1	2010-01-1200:06	8.3	2010-01-1223:46	7.9	84	69	2010-01-1218:14	1.6	310	8.4	320	2010-01-1212:50	7	
2010/1/13	1028.7	1029.5	1030.7	2010-01-1321:44	1026.5	2010-01-1315:07	10.9	16.3	2010-01-1314:09	7	2010-01-1305:39	3.8	64	40	2010-01-1312:43	2.3	70	9.7	30	2010-01-1316:44	0	
2010/1/14	1027.3	1028.1	1030.5	2010-01-1400:04	1023.9	2010-01-1414:57	13.8	19.3	2010-01-1414:11	9.2	2010-01-1400:25	5.3	57	41	2010-01-1413:48	3.7	70	12.8	60	2010-01-1410:19	0	
2010/1/15	1024.7	1025.5	1026.5	2010-01-1509:05	1022.6	2010-01-1514:27	17.3	22.6	2010-01-1514:03	12.4	2010-01-1500:28	8.8	58	40	2010-01-1512:40	3.7	70	11.2	70	2010-01-1513:46	0	
2010/1/16	1026	1026.8	1027.7	2010-01-1610:00	1024.1	2010-01-1603:04	16.8	18.3	2010-01-1608:19	15.4	2010-01-1611:00	12.3	75	67	2010-01-1600:01	4.4	60	15.2	60	2010-01-1613:35	0.2	
2010/1/17	1026.2	1027	1028.1	2010-01-1709:03	1024.6	2010-01-1714:10	17	18.5	2010-01-1712:43	16.1	2010-01-1702:40	12.3	74	72	2010-01-1707:33	4.1	70	14.4	60	2010-01-1716:47	0	
2010/1/18	1023	1023.8	1025.5	2010-01-1808:40	1020.7	2010-01-1814:57	20	24.6	2010-01-1814:06	16.4	2010-01-1801:13	13.2	66	47	2010-01-1814:05	3.6	70	11.3	60	2010-01-1811:41	0	
2010/1/19	1020.3	1021.1	1023	2010-01-1908:54	1018.4	2010-01-1915:12	21.7	27.1	2010-01-1912:08	16.6	2010-01-1904:28	14.3	63	46	2010-01-1912:03	2.1	90	8.6	100	2010-01-1917:16	0	
2010/1/20	1019.5	1020.3	1021.3	2010-01-2009:08	1017.7	2010-01-2014:56	22.4	27.8	2010-01-2012:00	18.2	2010-01-2005:37	15.6	66	51	2010-01-2012:30	1	310	7.6	310	2010-01-2013:14	0	
2010/1/21	1020	1020.8	1024.5	2010-01-2121:30	1017.3	2010-01-2113:04	20.4	26.9	2010-01-2111:24	15.6	2010-01-2123:58	15.2	73	56	2010-01-2111:23	3	70	14.9	60	2010-01-2114:19	0	
2010/1/22	1024.8	1025.6	1026.4	2010-01-2210:23	1023	2010-01-2203:04	14.9	16	2010-01-2211:43	14	2010-01-2222:46	12.7	86	81	2010-01-2200:59	4.7	70	13.8	90	2010-01-2210:44	4	
2010/1/23	1024.1	1024.9	1026.4	2010-01-2309:31	1022.1	2010-01-2314:49	14.3	14.9	2010-01-2318:56	13.4	2010-01-2306:38	12.6	89	86	2010-01-2300:15	4.9	70	13	60	2010-01-2314:42	2.2	
2010/1/24	1020.6	1021.4	1023.2	2010-01-2400:12	1017.7	2010-01-2415:36	17.6	20.5	2010-01-2415:32	14.9	2010-01-2400:01	16	91	84	2010-01-2415:55	2.5	70	10.3	60	2010-01-2401:46	3.5	
2010/1/25	1023.2	1024	1027.9	2010-01-2520:49	1018.2	2010-01-2501:50	15.7	18	2010-01-2510:19	13.4	2010-01-2523:14	13.7	88	77	2010-01-2522:16	2.3	70	12.4	60	2010-01-2516:53	18.8	
2010/1/26	1024.9	1025.7	1028.4	2010-01-2609:05	1021.7	2010-01-2623:46	16.1	17.8	2010-01-2613:19	13.7	2010-01-2600:01	10.9	71	66	2010-01-2612:54	5.3	70	16.7	60	2010-01-2611:38	0	
2010/1/27	1018.6	1019.4	1021.7	2010-01-2700:01	1016.5	2010-01-2714:55	19.8	24.7	2010-01-2712:04	14.9	2010-01-2705:10	14.8	73	59	2010-01-2712:01	1.5	70	9.1	80	2010-01-2700:10	0	
2010/1/28	1017.8	1018.6	1019.9	2010-01-2823:44	1016.4	2010-01-2814:17	19.1	21.8	2010-01-2811:26	17.5	2010-01-2806:30	15.9	82	72	2010-01-2810:40	1.6	150	8.2	80	2010-01-2822:19	1.1	
2010/1/29	1019.6	1020.4	1022.6	2010-01-2909:24	1017.6	2010-01-2915:55	19	21.6	2010-01-2913:50	16.9	2010-01-2906:42	14.8	77	65	2010-01-2914:30	4.3	70	12.3	60	2010-01-2913:29	0	
2010/1/30	1017.5	1018.3	1019.9	2010-01-3009:04	1015.3	2010-01-3014:01	19.8	26.3	2010-01-3014:02	17.4	2010-01-3007:50	15.7	78	56	2010-01-3013:54	1.8	60	9.9	300	2010-01-3014:33	0	
2010/1/31	1015.5	1016.2	1017.1	2010-01-3100:04	1013.5	2010-01-3123:57	21.5	27.1	2010-01-3113:30	16.9	2010-01-3106:41	16.5	74	53	2010-01-3113:05	0.9	280	6.6	300	2010-01-3115:04	0	
2010/2/1	1015.6	1016.4	1018.3	2010-02-0120:54	1013.3	2010-02-0102:35	18.7	22	2010-02-0110:52	15.7	2010-02-0123:38	15.4	81	75	2010-02-0120:51	3	70	14.8	60	2010-02-0117:41	0	
2010/2/2	1016.3	1017.1	1018.1	2010-02-0206:53	1013.7	2010-02-0214:14	17.9	20.8	2010-02-0213:42	15.6	2010-02-0200:08	14.5	81	72	2010-02-0212:34	3.9	70	13.8	60	2010-02-0212:16	T	
2010/2/3	1017.9	1018.7	1020	2010-02-0308:47	1016.4	2010-02-0314:21	17.1	18	2010-02-0313:35	15.7	2010-02-0307:44	14.8	86	81	2010-02-0300:52	2	70	9.3	50	2010-02-0313:02	2.1	
2010/2/4	1017.8	1018.6	1019.8	2010-02-0408:11	1015.4	2010-02-0413:58	17.8	20.5	2010-02-0412:49	16.2	2010-02-0407:26	15.4	86	78	2010-02-0413:25	1.7	270	10.2	70	2010-02-0412:15	2	
2010/2/5	1018	1018.8	1019.7	2010-02-0508:14	1015.8	2010-02-0514:13	18	21.4	2010-02-0512:20	16.5	2010-02-0504:37	15.5	86	76	2010-02-0511:42	1.1	240	7.1	320	2010-02-0514:47	T	
2010/2/6	1017.8	1018.6	1019.4	2010-02-0609:20	1015.8	2010-02-0613:15	18.2	19.8	2010-02-0611:15	17.3	2010-02-0606:37	16.1	88	80	2010-02-0613:50	1	220	4.5	300	2010-02-0607:15	7	
2010/2/7	1017.7	1018.5	1020	2010-02-0708:17	1015.6	2010-02-0716:26	17.1	18.4	2010-02-0716:29	15.8	2010-02-0707:48	15.6	91	86	2010-02-0716:18	1.4	300	10.5	310	2010-02-0706:06	9.7	
2010/2/8	1016	1016.8	1018.5	2010-02-0809:32	1014	2010-02-0814:07	20.5	26.7	2010-02-0813:40	16.2	2010-02-0805:35	16.9	80	64	2010-02-0812:40	1.4	80	7.8	60	2010-02-0810:36	0	
2010/2/9	1015.1	1015.9	1018.2	2010-02-0909:50	1012.6	2010-02-0914:28	21.2	26.8	2010-02-0914:44	17.7	2010-02-0905:26	17.5	80	61	2010-02-0915:08	0.9	240	6.2	290	2010-02-0915:29	0	
2010/2/10	1013.5	1014.3	1015.9	2010-02-1009:31	1010.7	2010-02-1015:22	22.8	29.8	2010-02-1013:34	18.5	2010-02-1004:44	18.7	79	52	2010-02-1013:33	0.8	150	8.1	330	2010-02-1017:40	0	
2010/2/11	1014.8	1015.5	1020.5	2010-02-1123:54	1012.4	2010-02-1114:40	23	29.5	2010-02-1113:25	15.6	2010-02-1123:49	17.9	74	55	2010-02-1113:01	2.5	70	17.9	70	2010-02-1120:30	T	
2010/2/12	1023.9	1024.7	1025.9	2010-02-1220:35	1020.1	2010-02-1200:18	12.6	15.7	2010-02-1200:01	11	2010-02-1218:52	9	79	73	2010-02-1216:01	5	70	15.6	70	2010-02-1201:03	T	
2010/2/13	1022	1022.8	1025.3	2010-02-1300:40	1016.7	2010-02-1323:58	13	15.5	2010-02-1323:27	11.5	2010-02-1300:15	11.3	90	83	2010-02-1300:05	4.5	70	11.9	90	2010-02-1309:00	21.3	
2010/2/14	1015.4	1016.2	1017.1	2010-02-1407:08	1013.4	2010-02-1413:58	17	22.7	2010-02-1413:24	14.1	2010-02-1404:21	14.6	86	68	2010-02-1412:43	1.5	260	9.8	70	2010-02-1400:15	0.2	
2010/2/15	1019.4	1020.2	1022.6	2010-02-1523:22	1016.8	2010-02-1500:01	14.4	16.4	2010-02-1512:25	12.8	2010-02-1523:59	12.6	89	79	2010-02-1513:15	2.2	70	10.9	70	2010-02-1512:02	15.4	
2010/2/16	1023.1	1023.9	1025.6	2010-02-1608:28	1021.1	2010-02-1601:50	11.9	13.1	2010-02-1601:01	11.3	2010-02-1606:35	10.7	92	88	2010-02-1614:59	2.6	60	10.3	60	2010-02-1614:48	41	

Data Preprocessing

- Drop data not needed & with missing values

4002	2020-12-16	1020.7	1024.4	1022.7	2020-12-1609:30	1019.3	2020-12-1615:22	15.5	16.2	2020-12-1611:16	...	1.5	2020-12-1613
4003	2020-12-17	1018.6	1022.3	1021.4	2020-12-1708:14	1016.0	2020-12-1723:15	16.6	19.5	2020-12-1713:56	...	0.5	2020-12-1700
4004	2020-12-18	1016.9	1020.5	1018.7	2020-12-1822:53	1015.0	2020-12-1814:17	18.9	21.3	2020-12-1814:30	...	0.5	2020-12-1814
4005	2020-12-19	1021.2	1024.8	1023.1	2020-12-1908:48	1018.4	2020-12-1900:16	16.8	18.7	2020-12-1900:01	...	T	2020-12-1908
4006	2020-12-20	...	...	...	...	...	...	...	...	...	...	...	
4007	2020-12-21	...	...	...	...	...	...	...	...	...	...	...	
4008	2020-12-22	...	...	...	...	...	...	...	...	...	...	...	
4009	2020-12-23	...	...	...	...	...	...	...	...	...	...	...	
4010	2020-12-24	...	...	...	...	...	...	...	...	...	...	...	
4011	2020-12-25	...	...	...	...	...	...	...	...	...	...	...	
4012	2020-12-26	...	...	...	...	...	...	...	...	...	...	...	
	2020-12-												

- Drop features that strongly relative to the target
- Drop features that are too professional or unreachable to the public
- Convert the data type into float so that can be trained with models

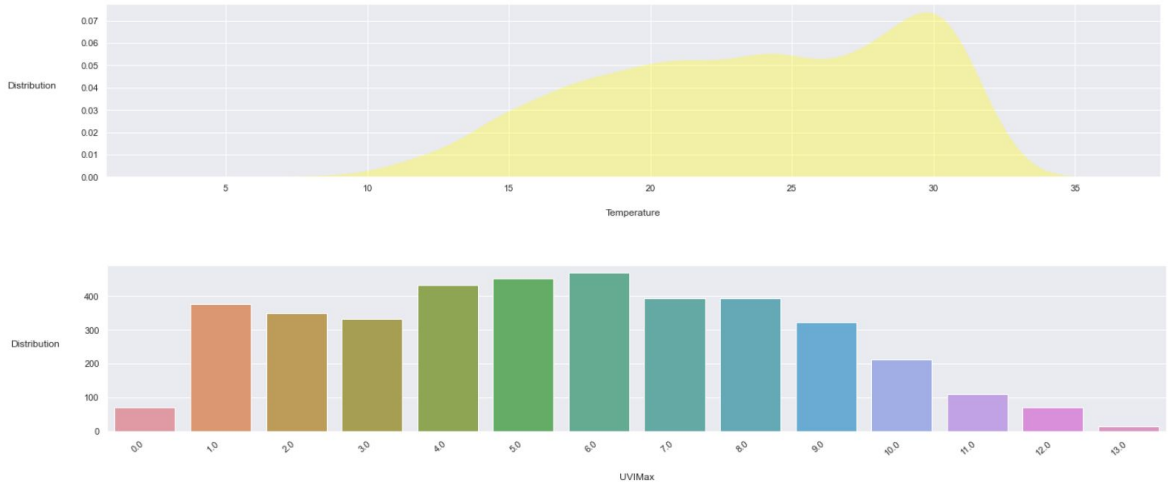
	StnPres	SeaPres	Temperature	WS	WD	SunShine	VisbMean	UVIMax	CloudAmount	PrecpHour
0	1018.2	1019.0	16.5	3.6	70	3.6	10.4	4	9.4	0.0
1	1016.4	1017.2	16.8	1.1	160	0.0	6.6	2	9.8	9.2
2	1016.8	1017.6	15.8	2.3	60	0.1	5.4	2	8.4	10.2
3	1015.3	1016.1	17.5	1.5	70	7.2	4.8	3	8.0	0.0
4	1018.9	1019.7	14.9	4.4	70	0.2	4.4	2	10.0	2.2
5	1020.2	1021.0	14.4	3.5	80	0.0	7.8	2	9.9	11.5

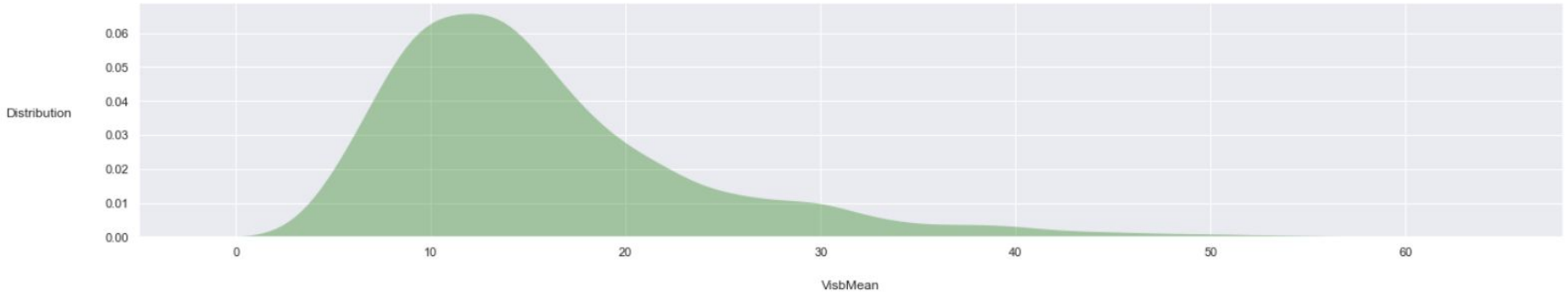
- Flush data so that the validation can have different results

	StnPres	SeaPres	Temperature	WS	WD	SunShine	VisbMean	UVIMax	CloudAmount	PrecpHour
3119	1001.0	1004.4	31.1	4.8	120	10.0	30.0	12	2.7	0.0
1396	1018.4	1019.2	22.9	2.6	80	7.8	16.9	6	3.4	0.0
972	1008.0	1008.7	27.5	1.3	180	1.9	7.4	6	8.6	5.1
1679	1002.0	1002.7	30.8	1.1	320	7.1	15.3	7	4.1	0.3
3081	997.3	1000.7	28.8	3.4	80	5.1	36.3	10	8.3	0.0
1838	1014.4	1017.9	20.3	3.5	80	7.1	16.8	5	7.9	2.0
2680	1008.3	1011.8	27.4	1.1	190	4.4	10.3	7	7.8	0.0
1524	1020.8	1021.6	14.6	2.5	80	0.0	9.8	1	10.0	6.2
328	1018.1	1018.9	18.4	1.5	180	0.0	6.8	0	9.6	12.7
848	1010.5	1011.2	23.3	0.8	170	1.5	11.1	7	8.5	8.4

Data Visualization

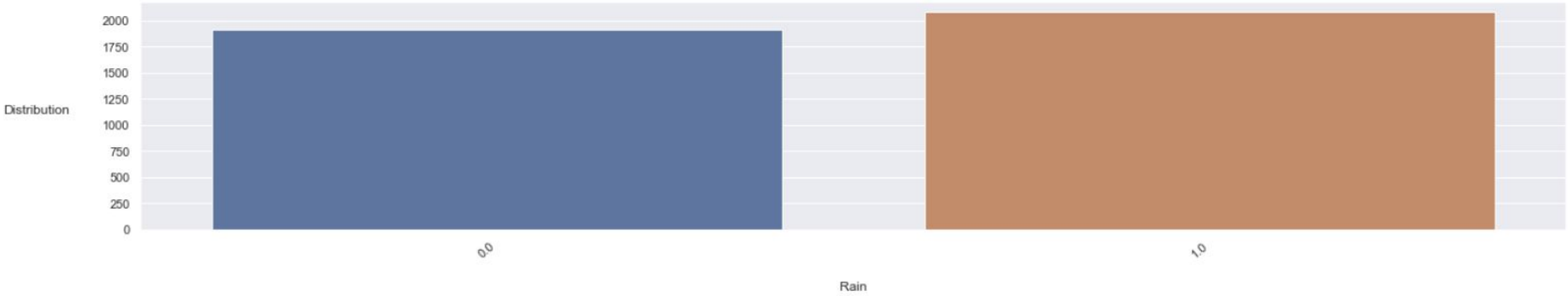
- Distribution of each feature
  - kde plot for continuous data
  - Histogram for discrete data




Distribution
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
50
60
VisbMean


Distribution
500
400
300
200
100
0
CloudAmount

Distribution of Target



Validation

- Holdout
  - Ratio: 7:3
- K-Fold
  - K = 3

Models

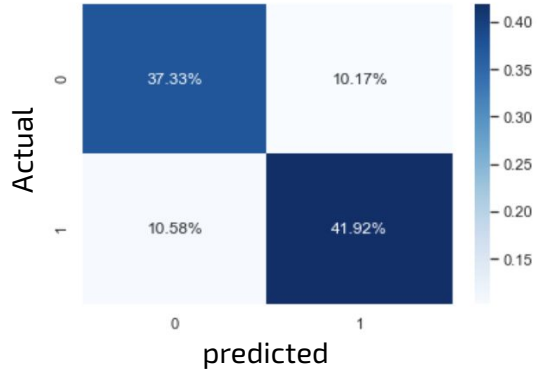
- Random Forest Classifier
- KNN Classifier
- Logistic Regression

Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier

rfclf_h = RandomForestClassifier(n_estimators=100)
rfclf_kf = RandomForestClassifier(n_estimators=100)
```

Results - Holdout

	precision	recall	f1-score	support
0.0	0.78	0.79	0.78	570
1.0	0.80	0.80	0.80	630
accuracy			0.79	1200
macro avg	0.79	0.79	0.79	1200
weighted avg	0.79	0.79	0.79	1200



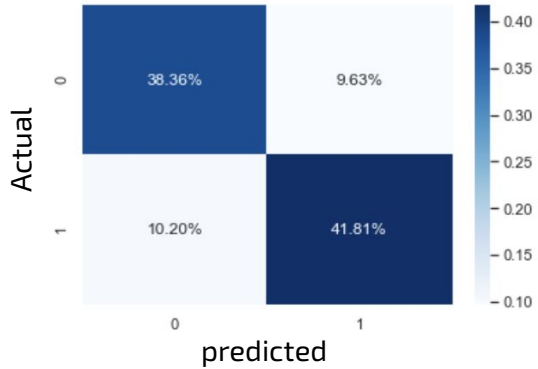
0: No rain
1: Rain

	Precision	Sensitivity
0	78%	79%
1	80%	80%

Accuracy: 79%

Results - K-Fold

	precision	recall	f1-score	support
0.0	0.79	0.80	0.79	1919
1.0	0.81	0.80	0.81	2080
accuracy			0.80	3999
macro avg	0.80	0.80	0.80	3999
weighted avg	0.80	0.80	0.80	3999



	Precision	Sensitivity
0	79%	80%
1	81%	80%

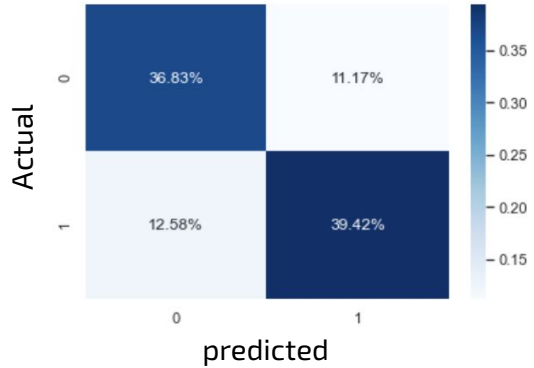
Accuracy: 80%

K-Nearest Neighbor Classifier

```
from sklearn.neighbors import KNeighborsClassifier

knnclf_h = KNeighborsClassifier(n_neighbors=20)
knnclf_kf = KNeighborsClassifier(n_neighbors=20)
```

Results - Holdout

	precision	recall	f1-score	support
0.0	0.75	0.77	0.76	576
1.0	0.78	0.76	0.77	624
accuracy			0.76	1200
macro avg	0.76	0.76	0.76	1200
weighted avg	0.76	0.76	0.76	1200

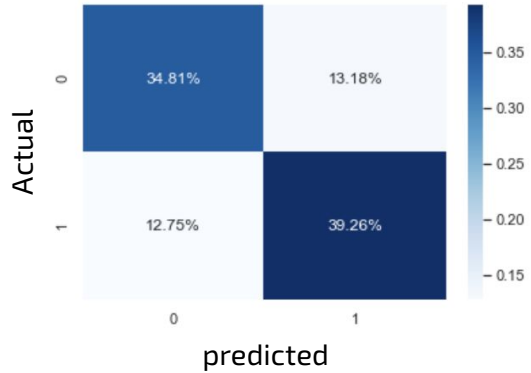


	Precision	Sensitivity
0	75%	77%
1	78%	76%

Accuracy: 76%

Results - K-Fold

	precision	recall	f1-score	support
0.0	0.73	0.73	0.73	1919
1.0	0.75	0.75	0.75	2080
accuracy			0.74	3999
macro avg	0.74	0.74	0.74	3999
weighted avg	0.74	0.74	0.74	3999



	Precision	Sensitivity
0	73%	73%
1	75%	75%

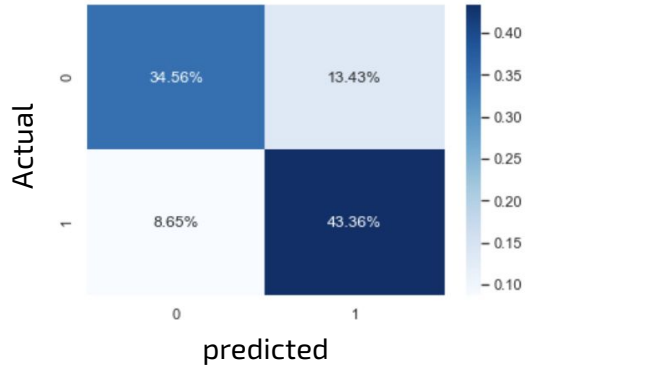
Accuracy: 74%

Logistic Regression

```
from sklearn.linear_model import LogisticRegression

log = LogisticRegression(max_iter=100000)
```

Results

	precision	recall	f1-score	support
0.0	0.80	0.72	0.76	1919
1.0	0.76	0.83	0.80	2080
accuracy			0.78	3999
macro avg	0.78	0.78	0.78	3999
weighted avg	0.78	0.78	0.78	3999

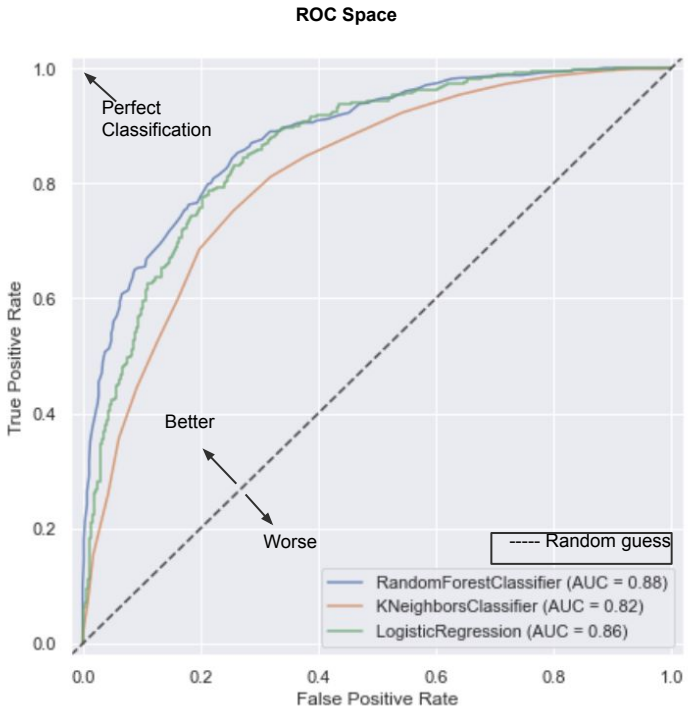


	Precision	Sensitivity
0	80%	72%
1	76%	83%

Accuracy: 78%

Results Visualization

- ROC curves
  - AUC: Area Under Curve



Application

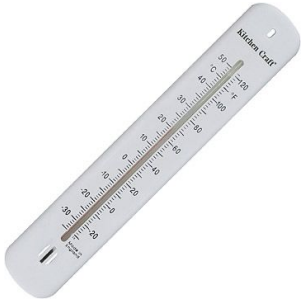
- Use known info to predict if it is raining
  - if not knowing some info → apply average value

```
StnPres: idk
1011.21
SeaPres: idk
1013.52
Temperature: 21
21.00
WS: idk
2.40
WD: idk
124.47
SunShine: 8
8.00
VisbMean: 10
10.00
UVIMax: idk
5.52
CloudAmount: 0
0.00

******************
**No rain today!**
******************
```

```
StnPres: idk
1011.21
SeaPres: idk
1013.52
Temperature: 21
21.00
WS: idk
2.40
WD: idk
124.47
SunShine: 8
8.00
VisbMean: 10
10.00
UVIMax: idk
5.52
CloudAmount: 10
10.00

****************************************
**Don't forget to bring your umbrella!**
****************************************
```

Vision

- We think we can combine the model with some detectors & IoT…
- design better UI…



THANKS
For
Listening

Q & A