data_engineering

March 6, 2018

0.1 Data Engineering

In [12]:
```
import pandas as pd
import numpy as np
```

In [13]:
```
measurements_df = pd.read_csv("Resources/hawaii_measurements.csv")
stations_df = pd.read_csv("Resources/hawaii_stations.csv")
```

In [14]:
```
measurements_df.head()
```

Out[14]:

	station	date	prcp	tobs
0	USC00519397	2010-01-01	0.08	65
1	USC00519397	2010-01-02	0.00	63
2	USC00519397	2010-01-03	0.00	74
3	USC00519397	2010-01-04	0.00	76
4	USC00519397	2010-01-06	NaN	73

In [15]:
```
stations_df.head()
```

Out[15]:

	station	name	latitude	longitude	elevation
0	USC00519397	WAIKIKI 717.2, HI US	21.2716	-157.8168	3.0
1	USC00513117	KANEOHE 838.1, HI US	21.4234	-157.8015	14.6
2	USC00514830	KUALOA RANCH HEADQUARTERS 886.9, HI US	21.5213	-157.8374	7.0
3	USC00517948	PEARL CITY, HI US	21.3934	-157.9751	11.9
4	USC00518838	UPPER WAHIAWA 874.3, HI US	21.4992	-158.0111	306.6

In [16]:
```
measurements_df[pd.isnull(measurements_df).any(axis=1)]
```

Out[16]:

	station	date	prcp	tobs
4	USC00519397	2010-01-06	NaN	73
26	USC00519397	2010-01-30	NaN	70

29	USC00519397	2010-02-03	NaN	67
43	USC00519397	2010-02-19	NaN	63
61	USC00519397	2010-03-11	NaN	73
72	USC00519397	2010-03-26	NaN	72
122	USC00519397	2010-05-21	NaN	77
176	USC00519397	2010-07-16	NaN	78
282	USC00519397	2010-11-04	NaN	73
294	USC00519397	2010-11-19	NaN	72
324	USC00519397	2010-12-26	NaN	74
341	USC00519397	2011-01-13	NaN	68
369	USC00519397	2011-02-12	NaN	68
390	USC00519397	2011-03-08	NaN	72
490	USC00519397	2011-06-24	NaN	77
586	USC00519397	2011-10-05	NaN	79
830	USC00519397	2012-06-08	NaN	77
831	USC00519397	2012-06-09	NaN	76
861	USC00519397	2012-07-09	NaN	77
901	USC00519397	2012-08-18	NaN	77
902	USC00519397	2012-08-19	NaN	76
1011	USC00519397	2012-12-06	NaN	69
1012	USC00519397	2012-12-07	NaN	69
1045	USC00519397	2013-01-10	NaN	72
1046	USC00519397	2013-01-11	NaN	72
1240	USC00519397	2013-07-24	NaN	79
1410	USC00519397	2014-01-10	NaN	72
1411	USC00519397	2014-01-11	NaN	70
1528	USC00519397	2014-05-08	NaN	73
1529	USC00519397	2014-05-09	NaN	77
...	...	...	...	...
19128	USC00516128	2016-06-05	NaN	73
19147	USC00516128	2016-06-25	NaN	73
19152	USC00516128	2016-07-01	NaN	74
19153	USC00516128	2016-07-04	NaN	74
19170	USC00516128	2016-07-23	NaN	74
19181	USC00516128	2016-08-03	NaN	74
19182	USC00516128	2016-08-04	NaN	74
19183	USC00516128	2016-08-05	NaN	75
19184	USC00516128	2016-08-06	NaN	77
19204	USC00516128	2016-08-27	NaN	74
19287	USC00516128	2016-11-20	NaN	74
19314	USC00516128	2016-12-18	NaN	67
19361	USC00516128	2017-02-04	NaN	66
19374	USC00516128	2017-02-18	NaN	72
19396	USC00516128	2017-03-13	NaN	69
19400	USC00516128	2017-03-18	NaN	70
19412	USC00516128	2017-03-31	NaN	76
19419	USC00516128	2017-04-08	NaN	76
19444	USC00516128	2017-05-04	NaN	74

```
19459  USC00516128  2017-05-20   NaN    70
19468  USC00516128  2017-05-30   NaN    72
19470  USC00516128  2017-06-03   NaN    74
19476  USC00516128  2017-06-10   NaN    72
19528  USC00516128  2017-08-01   NaN    72
19531  USC00516128  2017-08-05   NaN    77
19532  USC00516128  2017-08-06   NaN    79
19537  USC00516128  2017-08-11   NaN    72
19539  USC00516128  2017-08-13   NaN    80
19544  USC00516128  2017-08-18   NaN    76
19546  USC00516128  2017-08-20   NaN    78

[1447 rows x 4 columns]
```

```
In [17]: len(measurements_df)
```

```
Out[17]: 19550
```

About 7.5% of the precipition data is NaN, dates are missing, no temp data missing

```
In [18]: stations_df[pd.isnull(stations_df).any(axis=1)]
```

```
Out[18]: Empty DataFrame
         Columns: [station, name, latitude, longitude, elevation]
         Index: []
```

No data missing in station file

```
In [19]: measurements_df.to_csv("clean_measurements.csv", index = False)
```

```
In [20]: stations_df.to_csv("clean_stations.csv", index = False)
```