# wrangle_vs

February 10, 2016

## 1 Import data

```
In [1]: import pandas as pd
        f = pd.read_csv('../data/VSDATA_20150826.csv')
        f.head()
```

```
Out[1]:    NB_SCATS_SITE NM_REGION    QT_INTERVAL_COUNT  NB_DETECTOR   V00   V01  \
        0              2       DIO  2015-08-26 00:00:00            1 -1022 -1022
        1              2       DIO  2015-08-26 00:00:00            2 -1022 -1022
        2              2       DIO  2015-08-26 00:00:00            3 -1022 -1022
        3              2       DIO  2015-08-26 00:00:00            4 -1022 -1022
        4              2       DIO  2015-08-26 00:00:00            5 -1022 -1022

            V02   V03   V04   V05       ...          V89   V90   V91   V92   V93  \
        0 -1022 -1022 -1022 -1022       ...        -1022 -1022 -1022 -1022 -1022
        1 -1022 -1022 -1022 -1022       ...        -1022 -1022 -1022 -1022 -1022
        2 -1022 -1022 -1022 -1022       ...        -1022 -1022 -1022 -1022 -1022
        3 -1022 -1022 -1022 -1022       ...        -1022 -1022 -1022 -1022 -1022
        4 -1022 -1022 -1022 -1022       ...        -1022 -1022 -1022 -1022 -1022

            V94   V95  CT_RECORDS  QT_VOLUME_24HOUR  CT_ALARM_24HOUR
        0 -1022 -1022          96                 0               96
        1 -1022 -1022          96                 0               96
        2 -1022 -1022          96                 0               96
        3 -1022 -1022          96                 0               96
        4 -1022 -1022          96                 0               96

        [5 rows x 103 columns]
```

## 2 Filter data

Filter to site 2433 (mid-way along segment of Princes freeway monitored by bluetooth detector sites). Detectors 4-6 are in the outbound/westbound lanes.

```
In [2]: vols = f[(f["NB_SCATS_SITE"] == 2433) & f["NB_DETECTOR"].between(4,6)]
        vols
```

```
Out[2]:        NB_SCATS_SITE NM_REGION    QT_INTERVAL_COUNT  NB_DETECTOR  V00  V01  \
        26138           2433       GLI  2015-08-26 00:00:00            4   47   43
        26139           2433       GLI  2015-08-26 00:00:00            5   49   43
        26140           2433       GLI  2015-08-26 00:00:00            6   23   40

              V02  V03  V04  V05       ...          V89  V90  V91  V92  V93  V94  V95  \
```

```
26138   42   32   41   21       ...         104   73   78   67   81   60   54
26139   34   23   27   23       ...         100   77   74   71   75   63   56
26140   25   18    8   15       ...          89   81   73   65   62   45   32

       CT_RECORDS  QT_VOLUME_24HOUR  CT_ALARM_24HOUR
26138          96              9031                0
26139          96              9949                0
26140          96              9576                0

[3 rows x 103 columns]
```

# 3 Date range

Extract date from CSV data

```
In [3]: import datetime
        import pytz
        start_date = vols["QT_INTERVAL_COUNT"].iloc[0]
        start_date_parsed = datetime.datetime.strptime(start_date, '%Y-%m-%d 00:00:00')
        start_datetime = pytz.timezone('Australia/Melbourne').localize(start_date_parsed)
        date_range = pd.date_range(start_datetime, periods = 96, freq='15T')
        date_range[:10] # show first 10 rows
```

```
Out[3]: DatetimeIndex(['2015-08-26 00:00:00+10:00', '2015-08-26 00:15:00+10:00',
                       '2015-08-26 00:30:00+10:00', '2015-08-26 00:45:00+10:00',
                       '2015-08-26 01:00:00+10:00', '2015-08-26 01:15:00+10:00',
                       '2015-08-26 01:30:00+10:00', '2015-08-26 01:45:00+10:00',
                       '2015-08-26 02:00:00+10:00', '2015-08-26 02:15:00+10:00'],
                      dtype='datetime64[ns, Australia/Melbourne]', freq='15T')
```

# 4 Transform data

Transpose table. Label by time rather than interval. Use detector number as headers.

```
In [4]: dets = vols.T
        dets.columns = dets.loc["NB_DETECTOR"].values
        dets = dets.loc['V00':'V95']
        dets.index=date_range
        dets.head()
```

```
Out[4]:                            4   5   6
        2015-08-26 00:00:00+10:00  47  49  23
        2015-08-26 00:15:00+10:00  43  43  40
        2015-08-26 00:30:00+10:00  42  34  25
        2015-08-26 00:45:00+10:00  32  23  18
        2015-08-26 01:00:00+10:00  41  27   8
```

# 5 Export data

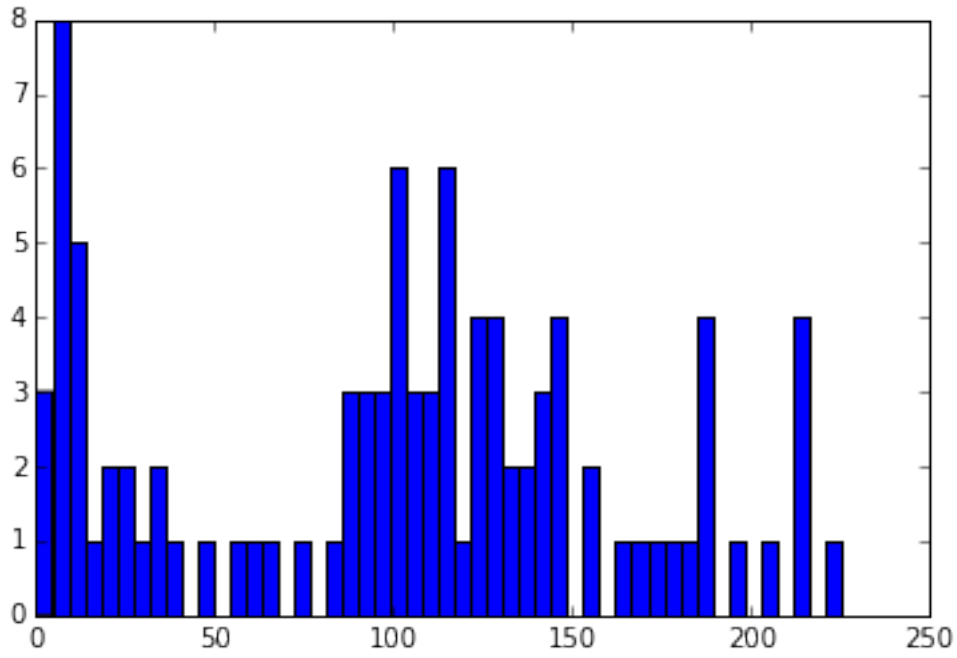
Extract just detector 6 (the rightmost lane)

```
In [5]: d6 = dets[6]
        d6.head()
```

```
Out[5]: 2015-08-26 00:00:00+10:00    23
        2015-08-26 00:15:00+10:00    40
        2015-08-26 00:30:00+10:00    25
        2015-08-26 00:45:00+10:00    18
        2015-08-26 01:00:00+10:00     8
        Freq: 15T, Name: 6, dtype: object
```

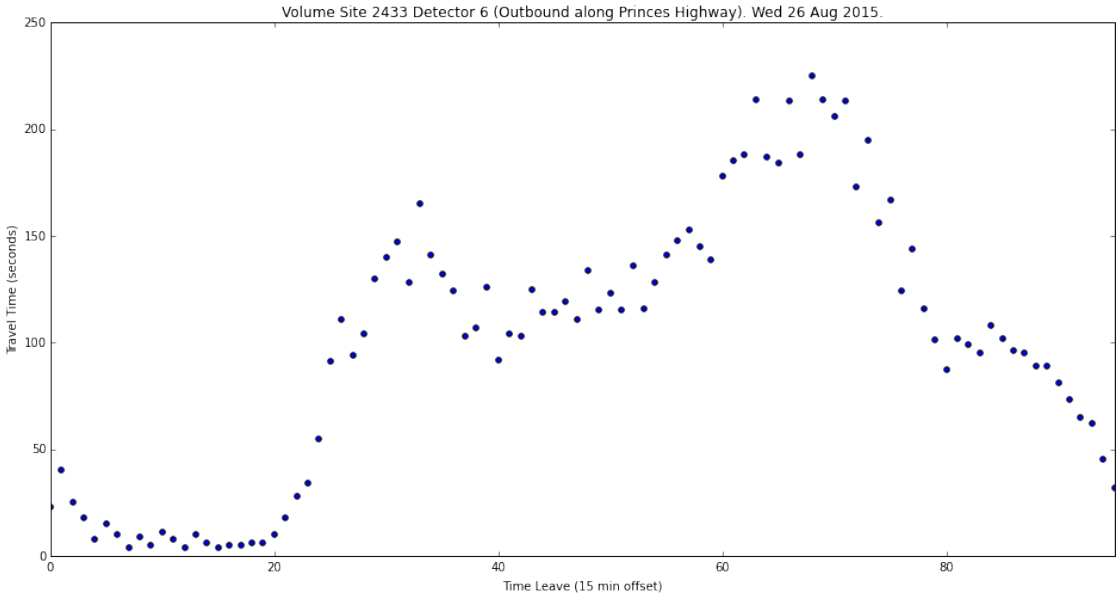
# 6 Plots

```
In [6]: import numpy as np
        %matplotlib inline
        import matplotlib.pyplot as plt
        bins = np.linspace(0, max(d6), 51)
        plt.hist(d6, bins=bins)
        plt.show()
```




```
In [7]: plt.figure(figsize=(16,8))
        plt.scatter(np.arange(len(d6)), d6.values)
        plt.title("Volume Site 2433 Detector 6 (Outbound along Princes Highway). Wed 26 Aug 2015.")
        plt.ylabel("Travel Time (seconds)")
        plt.xlabel("Time Leave (15 min offset)")
        plt.xlim([0,95])
        plt.ylim([0,None])
        plt.show()
```

```
/home/asimmons/anaconda3/envs/python2/lib/python2.7/site-packages/matplotlib/collections.py:590: FutureW
  if self._edgecolors == str('face'):
```


Volume Site 2433 Detector 6 (Outbound along Princes Highway). Wed 26 Aug 2015.
Travel Time (seconds)
Time Leave (15 min offset)
0
50
100
150
200
250
0
20
40
60
80