Homework 1 Problem 1

Overview

This exercise uses Python Pandas to analyze a dataset containing baseball statistics. The goal of the exercise is to determine if Oakland had a competitive advantage in certain years.

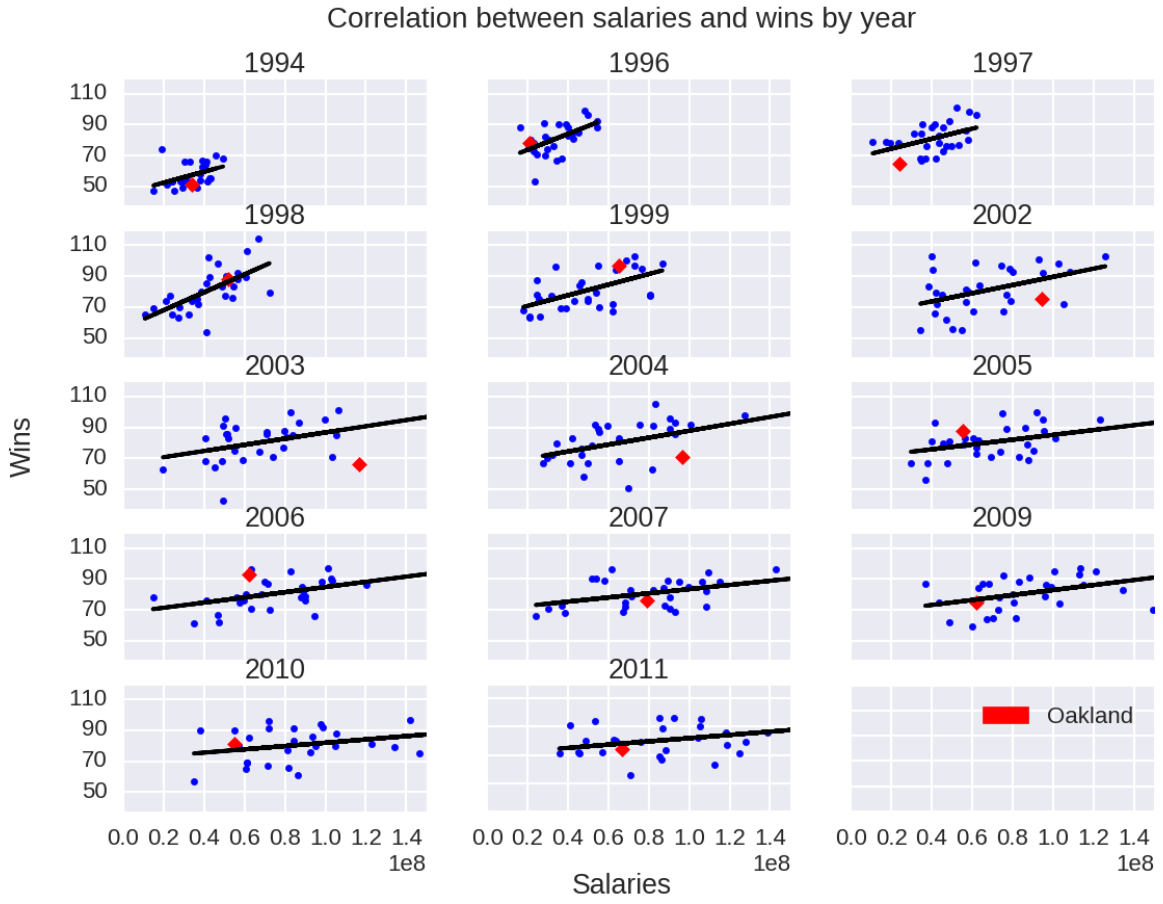
Useful commands

- **Check if a file exists**
  - import os
  - os.path.isfile('file')

- **Read in a csv file from the web into a pandas dataframe:** is the same as if the csv file was in your local folder
  - import pandas as pd
  - pd.read_csv('http://address.csv')

- **Subplots**
  - Share x and y axes (not putting axis labels on every subplot in the matrix)
    - fig,axes=plt.subplots(nrows,ncols,sharex=True,sharey=True)
  - Common x label
    - fig.text(0.5,0.01,'Common x label',ha='center')
  - Common y label
    - fig.text(0.04,0.5,'Common y label',va='center',rotation='vertical')
  - Common title
    - fig.suptitle('Common title')
      -

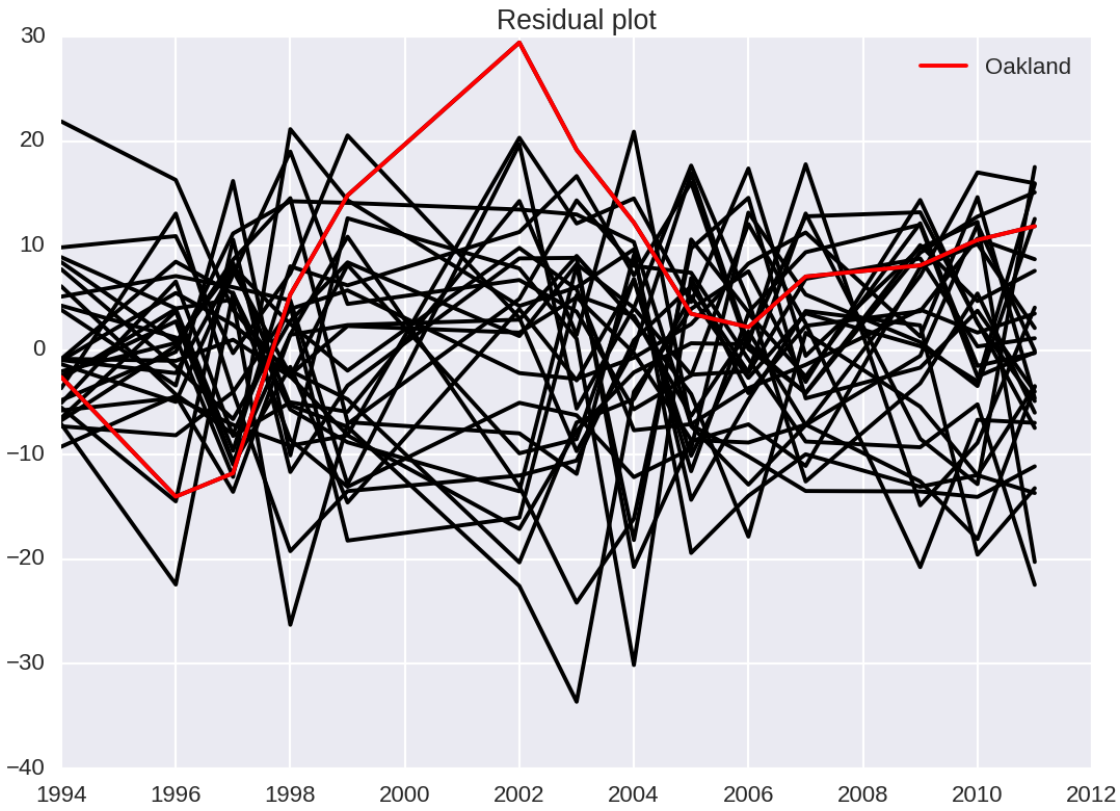
Figure 1:



Only certain years had statistically significant correlations between the total salary a team plays to its players and the wins for that team. Those years are included in the scatterplot in Figure 1, while the rest are excluded.

The red dot indicates how Oakland performed in each year. The figure shows that in certain years, Oakland paid a low salary while getting a large number of wins.

Figure 2:



A residual plot across years shows that Oakland had a large proportion of wins to salary approximately between the years 1999 and 2003.