**Esercitazione / simulazione esame**

**Nome del dataset:**

"PRSA\_data\_2010.1.1-2014.12.31.csv"

**Descrizione del dataset**:

Time period between Jan 1st, 2010 to Dec 31st, 2014. Missing data are denoted as NA

**No**: row number

**year**: year of data in this row

**month**: month of data in this row

**day**: day of data in this row

**hour**: hour of data in this row

**pm2.5**: outcome variable. PM2.5 concentration (ug/m^3)

**DEWP**: dew Point (â„ƒ)

**TEMP**: temperature (â„ƒ)

**PRES**: pressure (hPa)

**cbwd**: combined wind direction

**Iws**: cumulated wind speed (m/s)

**Is**: cumulated hours of snow

**Ir**: cumulated hours of rain

Lo scopo è quello di indagare quali siano i fattori atmosferici che influenzano le concentrazioni di pm2.5 nell'aria.

1) Produrre opportune **statistiche descrittive** che mostrino, a livello descrittivo, come il pm2.5 sia influenzato dalle diverse covariate.

2) Aggregare i dati a livello giornaliero utilizzando la funzione "media aritmetica" per **pm2.5 DEWP PRES**

e la funzione "massimo" per **Is Ir.** Utilizzare questo dataset per sviluppare i modelli successivi.

3) Analizzare ed eliminare eventuali osservazioni outlier prima di procedere con la costruzione del modello

4) Produrre un primo modello di regressione lineare verificando che siano rispettate tutte le ipotesi classiche del modello.

5) Migliorare il modello del punto 3 aggiungendo eventuali interazioni tra le variabili

6) **Multicollinearità** (covariate)

- matrice di correlazione

- conditional index

- vif

- tolleranza