Assignment	**1**		
Title	**Forest Fires Data Set**		
Data file	**01_forestfires.csv**		
Related file	**papers\01_fires.pdf**		
Sample size	**518**	**Number of variables**	**13**

This dataset is public available for research. The details are described in Cortez and Morais (2007). The dataset contains the following variables

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: "jan" to "dec"
4. day - day of the week: "mon" to "sun"
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m^2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84.

In this dataset we are interested to model the burned area of the forest as a function of the rest of the variables. Start with initial analysis (descriptive and pairwise comparisons) before proceeding to fit the appropriate model. Use transformations in order to make the model comply with the assumptions. Assess the goodness of fit (in sample) and the out of sample prediction using Root mean square error (RMSE) measures.

Reference

- Cortez P. and Morais A.. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimaraes, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. Available at: http://www.dsi.uminho.pt/~pcortez/fires.pdf .

Assignments	**2-3**		
Title	**Wine quality datasets**		
Data files	**02a_winequality-red.csv**		
	02b_winequality-white.csv		
Related file	**papers\02_wines.pdf**		
Sample size	**1600**	**Number of variables**	**12**

These two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, see at http://www.vinhoverde.pt/en/. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

The inputs include objective tests (e.g. PH values) and the output is based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

We are interested to understand what influences the quality of wines (and possibly predict how to create a good quality wine). The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods. For more information, read Cortez et al. (2009).

Available features:

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH
10. sulphates
11. alcohol
12. quality (score between 0 and 10)

Start using simple analysis including some descriptive analysis and pairwise comparisons and then proceed to the modeling of the wine quality. Identify good quality wines. What are the ingredients of a good quality wine. If you use this recipe, does this ensures a good quality wine?

[For both students with datasets 2a & 2b]: Compare the results for the red and white wine. What are the differences between the two different kind of wines in terms of good quality.

Source reference:

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Assignment	**4**
Title	**Parkinsons Telemonitoring Data Set**
Data files	**04_parkinsons_updrs.data.txt**
Related file	**papers\04_parkinson_1.pdf,04_parkinson_2.pdf**
Sample size	**5876** **Number of variables** **12**

This dataset is composed of a range of biomedical voice measurements from 42 people with early-stage Parkinson's disease recruited to a six-month trial of a telemonitoring device for remote symptom progression monitoring. The recordings were automatically captured in the patient's homes.

Columns in the table contain subject number, subject age, subject gender, time interval from baseline recruitment date, motor UPDRS, total UPDRS, and 16 biomedical voice measures. Each row corresponds to one of 5,875 voice recording from these individuals. The main aim of the data is to predict the motor and total UPDRS scores ('motor_UPDRS' and 'total_UPDRS') from the 16 voice measures.

The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around 200 recordings per patient, the subject number of the patient is identified in the first column. Further details about the biomedical voice measures can be found in Little et al. (2009).

The dataset was created by Athanasios Tsanas and Max Little of the University of Oxford, in collaboration with 10 medical centers in the US and Intel Corporation who developed the telemonitoring device to record the speech signals. The original study used a range of linear and nonlinear regression methods to predict the clinician's Parkinson's disease symptom score on the UPDRS scale.

ATTRIBUTE INFORMATION:

1. subject# - Integer that uniquely identifies each subject
2. age - Subject age
3. sex - Subject gender '0' - male, '1' - female
4. test_time - Time since recruitment into the trial. The integer part is the number of days since recruitment.
5. motor_UPDRS - Clinician's motor UPDRS score, linearly interpolated
6. total_UPDRS - Clinician's total UPDRS score, linearly interpolated
7. Jitter(%),Jitter(Abs),Jitter:RAP,Jitter:PPQ5,Jitter:DDP - Several measures of variation in fundamental frequency
8. Shimmer,Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,Shimmer:APQ11,Shimmer:DDA - Several measures of variation in amplitude
9. NHR, HNR - Two measures of ratio of noise to tonal components in the voice
10. RPDE - A nonlinear dynamical complexity measure
11. DFA - Signal fractal scaling exponent
12. PPE - A nonlinear measure of fundamental frequency variation

Analyze the above data starting from basic analysis and proceeding to regression models to predict motor and total UPDRS. Which variables are crucial for the prediction of these responses? Can the model be used for the prediction? How can we improve prediction?

Reference:

Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2009). 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering, 56(4):1015-1022

Source reference:

A Tsanas, MA Little, PE McSharry, LO Ramig (2010). 'Accurate telemonitoring of Parkinson.s disease progression by non-invasive speech tests', IEEE Transactions on Biomedical Engineering, **57** (4): 884 - 893.

Assignment	**5**		
Title	**Income, Education Level and Twins Data Set**		
Data files	**05_twins.dat.txt**		
Related file	**papers\ 05_twins.pdf**		
Sample size	**184**	**Number of variables**	**16**

The original study that used these data sought to answer what seems a simple question: *By how much will another year of schooling most likely raise one's income?* Attempts had been made to estimate the value of a year's education in previous studies, but previous estimates may have been imprecise for two reasons. The first, most obvious reason is the difficulty of extracting education's effect on income from the effect that other variables related to education have on income. That is, a worker's natural ability, his family background, and his innate intelligence are all possible confounding factors that must be controlled for to estimate the effect of education on income accurately. Thus this study interviewed twins, collecting information about education, income, and background. Because monozygotic twins (twins from a single egg) are genetically identical and have similar family backgrounds, they provide an excellent control for confounding variables.

The second difficulty in measuring the effect of income on education has to do with the false reporting of education levels, and this study is the first to address it. Since people are more likely to report a higher education level than they have actually attained, especially in face to face interviews, the data will contain a number of people with lower education levels in the higher education categories. Thus, since education usually increases income, estimates for the precise amount of this increase will be too low. To correct for this bias the researchers interviewed the twins separately and recorded two entries for each individual's education level: his self-reported education level and the education level reported by his twin. This allowed them to estimate the "measurement error" of reported education levels and correct for it. The result was a much higher estimate of the effect a year of education is likely to have on one's income. In fact, this study's estimates were higher than those of all previous studies, which did not correct for measurement error in education level. For more details, see in Ashenfelter & Krueger (1994).

The data were collected by a team of five interviewers at the 16th Annual Twins Day Festival in Twinsburg, Ohio, in August 1991. A booth was set up at the festival's main entrance, and an ad inviting all adult twins to participate in the survey was placed in the festival program. In addition, the interviews roamed the festival grounds, approaching all adult twins for an interview, and almost every pair of twins accepted. In total, 495 individuals over the age of 18 were interviewed.

The data file is comma delimited text. Each row contains information on a pair of twins for sixteen variables, which are explained below. Note that the each individual in a pair of twins was randomly assigned a number: twin 1 or twin 2. Missing data are indicated by a period.

Available variables:

1. DLHRWAGE.....the difference (twin 1 minus twin 2) in the logarithm of hourly wage, given in dollars.
2. DEDUC1.....the difference (twin 1 minus twin 2) in self-reported education, given in years.
3. AGE.....Age in years of twin 1.
4. AGESQ.....AGE squared.
5. HRWAGEH.....Hourly wage of twin 2.
6. WHITEH.....1 if twin 2 is white, 0 otherwise.
7. MALEH.....1 if twin 2 is male, 0 otherwise.
8. EDUCH.....Self-reported education (in years) of twin 2.
9. HRWAGEL.....Hourly wage of twin 1.
10. WHITEL.....1 if twin 1 is white, 0 otherwise.
11. MALEL.....1 if twin 1 is male, 0 otherwise.
12. EDUCL.....Self-reported education (in years) of twin 1.
13. DEDUC2.....the difference (twin 1 minus twin 2) in cross-reported education. Twin 1's cross-reported education, for example, is the number of years of schooling completed by twin 1 as reported by twin 2.
14. DTEN.....the difference (twin 1 minus twin 2) in tenure, or number of years at current job.
15. DMARRIED.....the difference (twin 1 minus twin 2) in marital status, where 1 signifies "married" and 0 signifies "unmarried".
16. DUNCOV.....the difference (twin 1 minus twin 2) in union coverage, where 1 signifies "covered" and 0 "uncovered".

Start from basic analysis of the data (descriptive statistics, graphs and pairwize comparisons) and proceed to regression models to identify the effect of the school years on the income. How can you embody the information from the twin in the model and how this reduces the bias?

Source reference:

Ashenfelter, O. and Krueger, A. (1994). "Estimates of the Economic Return to Schooling from a New Sample of Twins.", *The American Economic Review*, **84**, 1157-1173.

Assignments	**6-8**		
Title	**Minimum Wage, Unemployment and fast food Data Set**		
Data files	**06_fastfood.dat.txt**		
Related file	**papers\ 06_fastfood.pdf**		
Sample size	**411**	**Number of variables**	**43**

The accompanying data were used to study the effects of an increase in the minimum wage on unemployment. According to conventional economic theory, perfectly competitive employers will always cut their work force in response to any rise in the minimum wage. In practice, however, employer reactions are not so clear-cut: While some studies in the seventies confirmed the predictions of theory, earlier studies from the sixties, as well as more recent studies conducted in the early nineties, concluded that employment was unaffected by increases in the minimum wage.

This study sought to clarify the issue. The New Jersey minimum wage increase of April 1, 1992, which raised the minimum wage from $4.25 to $5.05, provided a perfect opportunity. Because the fast-food industry employs predominantly low-wage workers, because the absence of tips simplifies the measurement of wages, and because fast-food restaurants are relatively easy to sample, the study chose to assess the effects of the minimum wage increase on a random sample of Burger King, Wendys, KFC, and Roy Rogers restaurants in New Jersey and eastern Pennsylvania. The restaurants were interviewed about one month before, and about eight months after, the wage increase went into effect. Information was collected at each restaurant about variables such as the number of employees, various product prices, and store hours. This data provided everything necessary to calculate the impact that the minimum wage increase made not only on employment rates, but also on food prices, which can always be raised to compensate for higher wages. Finally, since unemployment is affected by factors other than the minimum wage, the study made careful use of controls. The data from the restaurants in Pennsylvania, as well as data on a number of restaurants in New Jersey that had been using the new five dollar minimum wage as a base salary even before the official increase went into effect, provided these naturally. More details from the study may be found in Card & Krueger (1994).

The final data set contains information on 410 restaurants randomly chosen from phonebooks in New Jersey and eastern Pennsylvania. For each restaurant information is provided on 46 variables; about half pertain to the period before the minimum wage increase, and about half concern the period after.

After it was chosen from the phonebook, each restaurant was called for a telephone survey. To elicit a response restaurants were called back as many as nine times, and the researchers obtained 410 completed interviews - an 87 percent response rate.

The data file is comma delimited text. The first row contains the list of variables and each remaining row contains the corresponding data for an individual restaurant. Missing data are indicated by a period. Explanations for all variable abbreviations are given below.

Available variables:

1. SHEET — sheet number (unique store id)
2. CHAIN — chain 1=Burger King; 2=KFC; 3=Roy Rogers; 4=Wendys
3. CO_OWNED — 1 if company owned
4. STATE — 1 if NJ; 0 if Pa
5. Dummies for location:
    a. SOUTHJ — 1 if in southern NJ
    b. CENTRALJ — 1 if in central NJ
    c. NORTHJ — 1 if in northern NJ
    d. PA1 — 1 if in PA, northeast suburbs of Phila
    e. PA2 — 1 if in PA, Easton etc
    f. SHORE — 1 if on NJ shore
6. First Interview:
    a. NCALLS — number of call-backs*
    b. EMPFT — # full-time employees
    c. EMPPT — # part-time employees
    d. NMGRS — # managers/ass't managers
    e. WAGE_ST — starting wage ($/hr)
    f. INCTIME — months to usual first raise
    g. FIRSTINC — usual amount of first raise ($/hr)
    h. BONUS — 1 if cash bounty for new workers
    i. PCTAFF — % employees affected by new minimum
    j. MEALS — free/reduced price code (See below)
    k. OPEN — hour of opening
    l. HRSOPEN — number hrs open per day
    m. PSODA — price of medium soda, including tax
    n. PFRY — price of small fries, including tax
    o. PENTREE — price of entree, including tax
    p. NREGS — number of cash registers in store
    q. NREGS11 — number of registers open at 11:00 am
7. Second Interview:
    a. TYPE2 — type 2nd interview 1=phone; 2=personal
    b. STATUS2 — status of second interview: see below
    c. DATE2 — date of second interview MMDDYY format
    d. NCALLS2 — number of call-backs*
    e. EMPFT2 — # full-time employees
    f. EMPPT2 — # part-time employees
    g. NMGRS2 — # managers/ass't managers
    h. WAGE_ST2 — starting wage ($/hr)
    i. INCTIME2 — months to usual first raise
    j. FIRSTIN2 — usual amount of first raise ($/hr)

k. SPECIAL2 1 if special program for new workers
l. MEALS2 free/reduced price code (See below)
m. OPEN2R hour of opening
n. HRSOPEN2 number hrs open per day
o. PSODA2 price of medium soda, including tax
p. PFRY2 price of small fries, including tax
q. PENTREE2 price of entree, including tax
r. NREGS2 number of cash registers in store
s. NREGS112 number of registers open at 11:00 am

8. Free/reduced Meal Variable:
    a. 0 = none
    b. 1 = free meals
    c. 2 = reduced price meals
    d. 3 = both free and reduced price meals
9. Second Interview Status:
    a. 0 = refused second interview (count = 1)
    b. 1 = answered 2nd interview (count = 399)
    c. 2 = closed for renovations (count = 2)
    d. 3 = closed "permanently" (count = 6)
    e. 4 = closed for highway construction (count = 1)
    f. 5 = closed due to Mall fire (count = 1)

*Note: number of call-backs = 0 if contacted on first call

[6a] Analyze the data of the first interview (also use the common variables in your analysis). Produce models to confirm the theory described above [if possible] and indentify the determinants of the size of fast foods, salaries and prices.

[6b] Analyze the data of the second interview. Produce models to confirm the theory described above [if possible] and indentify the determinants of the size of fast foods, salaries and prices.

[6c] Analyze the data of the differences between the two interviews. Produce models to indentify the theory described above [if possible] and the covariates that affect the size of fast foods, the salaries and the prices.

Source reference

Card, D. and Krueger, A. (1994). "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." .The American Economic Review 84, 772-793.

Assignments	**9**		
Title	**Beef demand Data Set**		
Data files	**09_BeefDemand**		
Sample size	**37**	**Number of variables**	**10**

The dataset consists of a several variables that may influence the demand for Beef in the United States. It provides an example of the influence of inflation in monetary time series data as well as providing some interesting statistical features in building demand models in regression.

The data set contains nominal prices, which are unadjusted for inflation. One way to adjust for the effects of inflation is to convert nominal prices into real prices by dividing the nominal price by the consumer price index for that year as a number. In so doing, the values are on a "constant ruler, namely in constant 1982-1984 dollars and cents. Economic theory would suggest a relationship between the beef consumption per capita and the inflation-adjusted beef price per pound with an increase in beef price per pound corresponding to a decrease in consumption.

Dataset variables:

1. Year: calendar year
2. ChickPrice: Chicken Retail Price in cents per pound
3. BeefPrice: Beef Retail Price in cents per pound
4. BeefConsump: Beef Consumption per capita in pounds
5. CPI: Consumer Price Index (CPI) for food
6. DPI: Disposable Personal Income per capita in dollars
7. RealChickPrice: Inflation-adjusted Chicken Retail Price in cents per pound
8. RealBeefPrice: Inflation-adjusted Beef Retail Price in cents per pound
9. RealDPI: Inflation-adjusted Disposable Personal Income per capita in dollars
10. (RDPI-Mean)^sq: The square of the difference between Inflation-adjusted Disposable Personal Income per capita and its mean

The ASCII file, BeefDemand.txt, is tab delimited. The first three of the last four variables are derived from their counterparts by dividing their values by the respective CPI and multiplying by 100. The last variable is derived by its description.

Assignment tasks: Analyze the above data in order to identify the determinants of beef consumption. Is your model confirming the economic theory. Start from the basics and then proceed to more advanced models.

SOURCES:

1. CPI source: http://www.ers.usda.gov/data/sdp/view.asp?f=livestock/89007/
2. Beef consumption per capita source: http://www.ers.usda.gov/Data/FoodConsumption/Spreadsheets/mtredsu.xls
3. Chicken retail price: http://www.ers.usda.gov/data/sdp/view.asp?f=livestock/89007/
4. Beef retail price source: http://www.ers.usda.gov/data/sdp/view.asp?f=livestock/94006/
5. 88~94 Disposable personal income pc: http://www.census.gov/prod/1/gen/95statab/income.pdf
6. 95~01 Disposable personal income pc: http://www.ccps.virginia.edu/.../statistical_abstract/Download_files/Section12download/12_10IncomeDisposable.xls

Assignments	**10**		
Title	London 2012 Olympics data		
Data files	**10_Olympics.csv**		
Sample size	**610**	**Number of variables**	**18**

This data set gives the names of the 203 participating countries as well as the number of gold, silver and bronze medals won by country, the total number of medals won by country, the Borda points by country, income per capita (in $10,000), population size (in 1,000,000,000), gross domestic product (GDP= income per capita multiplied by population size) and the polynomial variables of income per capita squared, population size squared, income per capita cubed, population size cubed, gross domestic product squared, gross domestic product cubed, natural log of income per capita, natural log of population size, and natural log of GDP.

A header line contains the names of the variables. The data are comma delimited. There are no missing values.

Available variables:

1. Country: The name of the country
2. GoldMedals: The number of gold medals won
3. Silver: The number of silver medals won
4. Bronze: The number of bronze medals won
5. TotalMedals: The number of total medals won
6. BordaPoints: The number of Borda points
7. Income: The income per capita (in $10,000)
8. PopnSize: The population size (in 1,000,000,000)
9. GDP: The gross domestic product (income multiplied by population size, units are in $10,000,000,000,000)
10. Income SQ: Income squared
11. PopnSQ: PopnSize squared
12. IncomeCubed: Income cubed
13. PopnCubed: PopnSize cubed
14. GDPSQ: GDP squared
15. GDPcubed: GDP cubed
16. Ln(Income): Natural log of Income
17. Ln(PopnSize): Natural log of PopnSize
18. Ln(GDP): Natural log of GDP

STORY BEHIND THE DATA:

While watching the London 2012 Olympics on television one of the authors was particularly impressed by the performance of Usain Bolt as he accumulated gold medals. From the small, relatively poor nation of Jamaica, he and his fellow Jamaicans were clearly dominating the sprinting events. While obviously, the strongest overall Olympic performances were from larger wealthier nations, Jamaican Olympic success was nevertheless spectacular. This raised the question: *Is it fair to compare Jamaica or any nation's success just on the number of gold or total medals and exactly how does population size and wealth influence a nation's performance*?

Assignment tasks: Analyze the above data starting from simple statistics and pairwise comparisons and moving to regression models. The aim is to identify which variables and factors influence the performance in the Olympic Games. Identify outliers and interpret them. Be careful on the selection of the response variable.

SOURCES: The data are publically available, and were obtained from http://espn.go.com/olympics/summer2012/medals and the list of nations at http://www.london2012.com/countries/. The data on income per capita and population size by country was obtained from Wikipedia and are available at http://www.csuchico.edu/math/theses-projects.shtml under Nathan Felton.

Assignments	**11**		
Title	**Greek super league data and betting odds**		
Data files	**11_super_league_2013-14.csv**		
Sample size	**306**	**Number of variables**	**53**

This dataset contains information for each football game in the Greek super league for season 2013 - 2014. All data are available in csv format, ready for use within a standard spreadsheet application. Please note that some abbreviations are no longer in use (in particular odds from specific bookmakers no longer used) and refer to data collected in earlier seasons. For a current list of what bookmakers are included in the dataset please visit http://www.football-data.co.uk/greecem.php.

Key to results data:
Div = League Division
Date = Match Date (dd/mm/yy)
HomeTeam = Home Team
AwayTeam = Away Team
FTHG = Full Time Home Team Goals
FTAG = Full Time Away Team Goals
FTR = Full Time Result (H=Home Win, D=Draw, A=Away Win)
HTHG = Half Time Home Team Goals
HTAG = Half Time Away Team Goals
HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win)

Match Statistics (where available)
Attendance = Crowd Attendance
Referee = Match Referee
HS = Home Team Shots
AS = Away Team Shots
HST = Home Team Shots on Target
AST = Away Team Shots on Target
HHW = Home Team Hit Woodwork
AHW = Away Team Hit Woodwork
HC = Home Team Corners
AC = Away Team Corners
HF = Home Team Fouls Committed
AF = Away Team Fouls Committed
HO = Home Team Offsides
AO = Away Team Offsides
HY = Home Team Yellow Cards
AY = Away Team Yellow Cards
HR = Home Team Red Cards
AR = Away Team Red Cards
HBP = Home Team Bookings Points (10 = yellow, 25 = red)
ABP = Away Team Bookings Points (10 = yellow, 25 = red)

Bb1X2 = Number of BetBrain bookmakers used to calculate match odds averages and maximums

Key to 1X2 (match) betting odds data:
B365H = Bet365 home win odds
B365D = Bet365 draw odds
B365A = Bet365 away win odds
BSH = Blue Square home win odds
BSD = Blue Square draw odds
BSA = Blue Square away win odds
BWH = Bet&Win home win odds
BWD = Bet&Win draw odds
BWA = Bet&Win away win odds
GBH = Gamebookers home win odds
GBD = Gamebookers draw odds
GBA = Gamebookers away win odds
IWH = Interwetten home win odds
IWD = Interwetten draw odds
IWA = Interwetten away win odds
LBH = Ladbrokes home win odds
LBD = Ladbrokes draw odds
LBA = Ladbrokes away win odds
PSH = Pinnacle Sports home win odds
PSD = Pinnacle Sports draw odds
PSA = Pinnacle Sports away win odds
SOH = Sporting Odds home win odds
SOD = Sporting Odds draw odds
SOA = Sporting Odds away win odds
SBH = Sportingbet home win odds
SBD = Sportingbet draw odds
SBA = Sportingbet away win odds
SJH = Stan James home win odds
SJD = Stan James draw odds
SJA = Stan James away win odds
SYH = Stanleybet home win odds
SYD = Stanleybet draw odds
SYA = Stanleybet away win odds
VCH = VC Bet home win odds
VCD = VC Bet draw odds
VCA = VC Bet away win odds
WHH = William Hill home win odds
WHD = William Hill draw odds
WHA = William Hill away win odds

Key to Asian handicap betting odds:
BbAH = Number of BetBrain bookmakers used to

BbMxH = Betbrain maximum home win odds
BbAvH = Betbrain average home win odds
BbMxD = Betbrain maximum draw odds
BbAvD = Betbrain average draw win odds
BbMxA = Betbrain maximum away win odds
BbAvA = Betbrain average away win odds

Key to total goals betting odds:
BbOU = Number of BetBrain bookmakers used to calculate over/under 2.5 goals (total goals) averages and maximums
BbMx>2.5 = Betbrain maximum over 2.5 goals
BbAv>2.5 = Betbrain average over 2.5 goals
BbMx<2.5 = Betbrain maximum under 2.5 goals
BbAv<2.5 = Betbrain average under 2.5 goals

GB>2.5 = Gamebookers over 2.5 goals
GB<2.5 = Gamebookers under 2.5 goals
B365>2.5 = Bet365 over 2.5 goals
B365<2.5 = Bet365 under 2.5 goals

Asian handicap averages and maximums
BbAHh = Betbrain size of handicap (home team)
BbMxAHH = Betbrain maximum Asian handicap home team odds
BbAvAHH = Betbrain average Asian handicap home team odds
BbMxAHA = Betbrain maximum Asian handicap away team odds
BbAvAHA = Betbrain average Asian handicap away team odds

GBAHH = Gamebookers Asian handicap home team odds
GBAHA = Gamebookers Asian handicap away team odds
GBAH = Gamebookers size of handicap (home team)
LBAHH = Ladbrokes Asian handicap home team odds
LBAHA = Ladbrokes Asian handicap away team odds
LBAH = Ladbrokes size of handicap (home team)
B365AHH = Bet365 Asian handicap home team odds
B365AHA = Bet365 Asian handicap away team odds
B365AH = Bet365 size of handicap (home team)

Assignment targets: Analyze the above data starting from simple descriptive statistics and moving forward to pairwize comparisons. Try to see if the information between all bookies is similar. Select one to proceed with a predictive model. Can you find a model which will accurately predict the final result or at least the odds of the bookie you have selected? Interpret the results.

Additional optional task: Try to estimate probabilities of each position using simulation.

Assignments	**12**		
Title	FRAMINGHAM Study dataset – 2		
Data files	**12_framdata.sav**		
Sample size	**1406**	**Number of variables**	**6**

The data for this assignment come from one of the most famous cohort studies performed in patients with cardiovascular disease, the Farmingham Heart Study. The study began in 1948 and is still ongoing, with the study's database being continuously updated. File farmdata.sav includes measurements in 1406 patients suffering from cardiovascular disease. Specifically, the variables included in the dataset are:

1. AGE: Patients' age
2. SBP: Systolic Blood Pressure
3. DBP: Diastolic Blood Pressure
4. CHOL: Total Cholesterol level
5. CIG: Number of cigarettes smoked per day
6. MALE: Gender, 0 for Females, 1 for Males

You are asked to answer the following questions by analyzing the data

a. Present suitable descriptive measures for all the variables individually, as well as for their (pairwise) associations

b. Is there a correlation between systolic and diastolic blood pressure?

c. Is the number of cigarettes associated with the patients' systolic pressure?

d. Estimate a linear model that examines the association of cholesterol levels with the rest of the variables in the dataset

Source of the Data:

- R. Dawber, M.D., Gilcin F. Meadors, M.D., M.P.H., and Felix E. Moore, Jr., *National Heart Institute, National Institues of Health, Public Health Service, Federal Security Agensy, Washington, D. C., Epidemiological Approaches to Heart Disease: The Framingham Study* Presented at a Joint Session of the Epidemiology, Health Officers, Medical Care, and Statistics Sections of the American Public Health Association, at the Seventy-eighth Annual Meeting in St. Louis, Mo., November 3, 1950.
- Daniel Levy and Susan Brink. (2005). *A Change of Heart: How the People of Framingham, Massachusetts, Helped Unravel the Mysteries of Cardiovascular Disease.* Knopf. ISBN 0-375-41275-1.

Assignments	**13**		
Title	**Supermarket spent dataset**		
Data files	**13_supermarkets.sav**		
Sample size	**637**	**Number of variables**	**9**

The source of the data of this assignment is a survey which conducted in order to identify the factors which influence the annual amount of money spent by supermarket clients. For the management of multi-stores and super markets, the aim is to make appropriate changes to increase this amount of money spent in their stores. A list of the available variables follows:

- Shop: The shop that each consumer usually buys everyday products. This is a categorical variable with codes from 1 to 18 which correspond to 18 large supermarket chains.
- Spen_yr: The amount spend annually in $.
- Car: Whether he uses his car to approach the store (0=NO, 1=YES)
- Prices: Evaluation of the price level by the customer (1=very high, 2=high, 3=medium prices, 4= low, 5= very low).
- Qfresh: Evaluation of the quality of fresh products (1=very low, 2= low, 3= medium, 4= high, 5= very high).
- Qpack: Evaluation of the quality of packed products (1=very low, 2= low, 3= medium, 4= high, 5= very high).
- Queues: Evaluation of the length of the queue in the counters (1=very long, 2= long, 3= medium length queues, 4= short, 5= very short).
- Distance: Evaluation of the distance from his house (1=very long, 2= long, 3= medium length distance, 4= small distance, 5= very small distance).
- Sex: Consumer's gender (0=male, 1=female)

Present the appropriate descriptive measures and study the pairwise associations. Are there any differences in the amount of money spent between males and females or according to his price evaluation? Fit a statistical model which studies the behavior of the annual customer spent. Which are the actions that the management of a supermarket should consider in order to increase the profit?

Assignments	**14-15**		
Title	**Student Performance Data Set**		
Data files	**14-15_student.zip**		
Sample size	**649**	**Number of variables**	**33**

This dataset approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details)

Attribute Information:

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)

16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

References

- P. Cortez and A. Silva (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)* pp. 5-12, Porto, Portugal, EUROSIS, ISBN 978-9077381-39-7.

  Available at: http://www3.dsi.uminho.pt/pcortez/student.pdf .

- Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science..

RESEARCH TASK FOR ASSIGMENT 14

Analyze the above data and setup a model to predict the final and the intermediate grades for Maths.

RESEARCH TASK FOR ASSIGMENT 15

Analyze the above data and setup a model to predict the final and the intermediate grades for Portuguese.

RESEARCH TASKS FOR BOTH ASSIGMENTS 14 & 15

a) Perform an exploratory analysis for your dataset.
b) Use visual methods to describe the most important findings.
c) Study the pairwise associations between the variables
d) Construct regression models to assess your responses.
e) **Both Assignments (with collaboration)**: Compare the two models. What are the differences between the two subjects?

Assignments	**16-17**		
Title	**Football Data**		
Data files	**Collect your own data**		
Sample size	**?**	**Number of variables**	**?**

Assignment 16

Collect the football data of the previous season for the Greek League (or any other Major European League) and do the following:

a) Construct a model for the prediction of the final score.

b) Evaluate the fit of your model.

c) Use simulation techniques to reproduce the league and calculate the average points, the expected ranking and the probability of each place in the league.

Assignment 17

On the same datasets, focus only on the final score (win, draw, loss) and perform the same analysis for (a-c)

Both assignments 16 & 17: Compare the performance of the two models. Which do you prefer and why?

Assignments	**18**		
Title	**Computer Hardware Data Set**		
Data files	**18_machine.data.txt; 18_machine.names.txt**		
Sample size	**209**	**Number of variables**	**9**

Data Set Information:

The estimated relative performance values were estimated by the authors using a linear regression method. See their article (pp 308-313) for more details on how the relative performance values were set.

Attribute Information

1. Vendor name: 30 (adviser, amdahl,apollo, basf, bti, burroughs, c.r.d, cambex, cdc, dec, dg, formation, four-phase, gould, honeywell, hp, ibm, ipl, magnuson, microdata, nas, ncr, nixdorf, perkin-elmer, prime, siemens, sperry, sratus, wang)
2. Model Name: many unique symbols
3. MYCT: machine cycle time in nanoseconds (integer)
4. MMIN: minimum main memory in kilobytes (integer)
5. MMAX: maximum main memory in kilobytes (integer)
6. CACH: cache memory in kilobytes (integer)
7. CHMIN: minimum channels in units (integer)
8. CHMAX: maximum channels in units (integer)
9. PRP: published relative performance (integer)
10. ERP: estimated relative performance from the original article (integer)

ASSIGNMENT TASKS

a. Perform an exploratory analysis for your dataset.
b. Use visual methods to describe the most important findings.
c. Study the pairwise associations between the variables
d. Construct regression models to assess PRP.
e. Find and confirm the regression model from which ERP was calculated.

Source

- Creator: Phillip Ein-Dor and Jacob Feldmesser, Faculty of Management, Tel Aviv University; Israel.
- Donor: David W. Aha (aha '@' ics.uci.edu)

References

- Kibler,D. & Aha,D. (1988). Instance-Based Prediction of Real-Valued Attributes. In Proceedings of the CSCSI (Canadian AI) Conference.
- Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science..

Assignments	**19**		
Title	**Energy efficiency Data Set**		
Data files	**19_ENB2012_data.xlsx**		
Sample size	**768**	**Number of variables**	**10**

Data Set Information

We perform energy analysis using 12 different building shapes simulated in Ecotect. The buildings differ with respect to the glazing area, the glazing area distribution, and the orientation, amongst other parameters. We simulate various settings as functions of the afore-mentioned characteristics to obtain 768 building shapes. The dataset comprises 768 samples and 8 features, aiming to predict two real valued responses (y1 and y2).

Attribute Information

The dataset contains eight attributes (or features, denoted by X1...X8) and two responses (or outcomes, denoted by y1 and y2). The aim is to use the eight features to predict each of the two responses.

- X1 - Relative Compactness - No units
- X2 - Surface Area - m²
- X3 - Wall Area - m²
- X4 - Roof Area - m²
- X5 - Height - m
- X6 - Orientation - 2:North, 3:East, 4:South, 5:West - No units
- X7 - Glazing Area - 0%, 10%, 25%, 40% (of floor area) - No units
- X8 - Glazing Variations - 1:Uniform, 2:North, 3:East, 4:South, 5:West
- Y1 - Heating Load - kWh/m²
- Y2 - Cooling Load - kWh/m²

ASSIGNMENT TASKS

f) Perform an exploratory analysis for your dataset.

g) Use visual methods to describe the most important findings.

h) Study the pairwise associations between the variables

i) Construct regression models to assess y1 and y2.

Source

The dataset was created by Angeliki Xifara (angxifara '@' gmail.com, Civil/Structural Engineer) and was processed by Athanasios Tsanas (tsanasthanasis '@' gmail.com, Oxford Centre for Industrial and Applied Mathematics, University of Oxford, UK).

References

- Tsanas A. and Xifara A. (2012): Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools,

Energy and Buildings, Vol. **49**, pp. 560-567. (the paper can be accessed at http://people.maths.ox.ac.uk/tsanas/Preprints/ENB2012.pdf)

- Tsanas A. (2012). *Accurate telemonitoring of Parkinson disease symptom severity using nonlinear speech signal processing and statistical machine learning*, D.Phil. thesis, University of Oxford, 2012 (For further details on the data analysis methodology; available at [Web Link])
- Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Assignments	**20**		
Title	**Yacht Hydrodynamics Data Set**		
Data files	**20_yacht_hydrodynamics.data.txt**		
Sample size	**308**	**Number of variables**	**7**

Data Set Information

Prediction of residuary resistance of sailing yachts at the initial design stage is of a great value for evaluating the ship performance and for estimating the required propulsive power. Essential inputs include the basic hull dimensions and the boat velocity.

The Delft data set comprises 308 full-scale experiments, which were performed at the Delft Ship Hydromechanics Laboratory for that purpose. These experiments include 22 different hull forms, derived from a parent form closely related to the Standfast designed by Frans Maas.

Attribute Information:

Variations concern hull geometry coefficients and the Froude number:

1. Longitudinal position of the center of buoyancy.
2. Prismatic coefficient.
3. Length-displacement ratio.
4. Beam-draught ratio.
5. Length-beam ratio.
6. Froude number.

The measured variable is the residuary resistance per unit weight of displacement:

7. Residuary resistance per unit weight of displacement.

ASSIGNMENT TASKS

f. Perform an exploratory analysis for your dataset.

g. Use visual methods to describe the most important findings.

h. Study the pairwise associations between the variables

i. Construct a regression model (or more) to assess the residuary resistance per unit weight of displacement.

Source:

- Creator: Ship Hydromechanics Laboratory, Maritime and Transport Technology Department, Technical University of Delft.
- Donor: Dr Roberto Lopez E-mail: roberto-lopez '@' users.sourceforge.net

References

- Gerritsma J., Onnink, R. and Versluis A. (1981). Geometry, resistance and stability of the delft systematic yacht hull series. *In International Shipbuilding Progress*, volume **28**, pages 276-297.
- Ortigosa I., Lopez R. and Garcia J. (2007). A neural networks approach to residuary resistance of sailing yachts prediction. In *Proceedings of the International Conference on Marine Engineering MARINE 2007*.
- Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Assignments	**21-22**		
Title	**Exploring Relationships in Body Dimensions**		
Data files	**21_body.dat.txt**		
Sample size	507	**Number of variables**	25

This dataset contains 21 body dimension measurements as well as age, weight, height, and gender on 507 individuals. The 247 men and 260 women were primarily individuals in their twenties and thirties, with a scattering of older men and women, all exercising several hours a week.

STORY BEHIND THE DATA

The first two authors investigated the correspondence between frame size, girths, and weight of physically active young men and women, most of whom were within normal weight range. One goal of this investigation was to develop predictive techniques to assess the lean/fat body composition of individuals.

SOURCE: Measurements were initially taken by the first two authors - Grete Heinz and Louis J. Peterson - at San Jose State University and at the U.S. Naval Postgraduate School in Monterey, California. Later, measurements were taken at dozens of California health and fitness clubs by technicians under the supervision of one of these authors.

VARIABLE DESCRIPTIONS:

- **Skeletal Measurements:**
  - ✓ 1 - 4 Biacromial diameter (see Fig. 2)
  - ✓ 6 - 9 Biiliac diameter, or "pelvic breadth" (see Fig. 2)
  - ✓ 11 - 14 Bitrochanteric diameter (see Fig. 2)
  - ✓ 16 - 19 Chest depth between spine and sternum at nipple level,
  - ✓ mid-expiration
  - ✓ 21 - 24 Chest diameter at nipple level, mid-expiration
  - ✓ 26 - 29 Elbow diameter, sum of two elbows
  - ✓ 31 - 34 Wrist diameter, sum of two wrists
  - ✓ 36 - 39 Knee diameter, sum of two knees
  - ✓ 41 - 44 Ankle diameter, sum of two ankles
- **Girth Measurements:**
  - 46 - 50 Shoulder girth over deltoid muscles
  - 52 - 56 Chest girth, nipple line in males and just above breast
  - tissue in females, mid-expiration
  - 58 - 62 Waist girth, narrowest part of torso below the rib cage,
  - average of contracted and relaxed position
  - 64 - 68 Navel (or "Abdominal") girth at umbilicus and iliac crest,
  - iliac crest as a landmark
  - 70 - 74 Hip girth at level of bitrochanteric diameter
  - 76 - 79 Thigh girth below gluteal fold, average of right and left
  - girths

- 81 - 84 Bicep girth, flexed, average of right and left girths
- 86 - 89 Forearm girth, extended, palm up, average of right and
- left girths
- 91 - 94 Knee girth over patella, slightly flexed position, average
- of right and left girths
- 96 - 99 Calf maximum girth, average of right and left girths
- 101 -104 Ankle minimum girth, average of right and left girths
- 106 -109 Wrist minimum girth, average of right and left girths

- **Other Measurements:**

  - 111-114 Age (years)
  - 116-120 Weight (kg)
  - 122-126 Height (cm)
  - 128 Gender (1 - male, 0 - female)

The first 21 variables are all measured in centimeters (cm).

Values are separated by blanks. There are no missing values.

SUBMITTED BY:

Grete Heinz
24710 Upper Trail
Carmel, CA 93923
USA
goguh@aol.com

Louis J. Peterson
Department of Health Sciences
San Josı State University
One Washington Square
San Josı, California 95192
USA

Roger W. Johnson
Department of Mathematics and Computer Science
South Dakota School of Mines and Technology
501 East St. Joseph Street
Rapid City, South Dakota 57701
USA
Roger.Johnson@sdsmt.edu

Carter J. Kerk
Industrial Engineering Program
South Dakota School of Mines and Technology
501 East St. Joseph Street
Rapid City, South Dakota 57701
USA
Carter.Kerk@sdsmt.edu

--

ASSIGNMENT 21

Use Skeletal and other measurements in your assignment.

a) Analyse the data using descriptive and visual methods.

b) Report pairwise comparisons.

c) Contract two models. One for the weight and one for the height using skeletal and other measurements.

ASSIGNMENT 22

Use Girth and other measurements in your assignment.

a) Analyse the data using descriptive and visual methods.

b) Report pairwise comparisons, the associated inference and interpretation.

c) Contract two models. One for the weight and one for the height using girth and other measurements. Check for the assumptions and perform diagnostic tests. Interpret the final results.

BOTH ASSIGNMENTS 21 & 22 (in collaboration)

d) Compare the regression models you have constructed. Which set of variables is better for predicting weight and height.

e) Construct a model with all covariates for weight. Do we improve our predictive performance?

Assignments	**23**		
Title	**Four-Mile Run Dataset**		
Data files	**23_Four-Mile_Run_Dataset.csv**		
Sample size	19	**Number of variables**	14

After losing a couple friends to smoking-related cancer at relatively young ages, Kevin decided to make some positive changes in his own life at the age of 47. Not only did he commit to give up his casual cigarette smoking, but he also decided to begin a running program. Up to this point in his life, Kevin had never really exercised regularly and he was approximately 55 pounds overweight. He began his running program slowly, intermittently jogging slowly and walking around his neighborhood. After a period of time, he was able to run for relatively short distances and he eventually built up his stamina enough to run a four-mile loop near his home. In order to track his progress, Kevin decided to purchase a GPS watch that would monitor his training data, including his heart rate, running pace, and calories burned, among other variables. After collecting training data for nineteen different runs, Kevin wants to analyze the data in order to see how he is progressing in his exercise program.

Specifically, Kevin is interested in determining how effective his training regimen is at improving his cardiovascular fitness and how he might modify his effort on individual runs in order to optimize overall health benefits.

The data that appear in this dataset were collected by a Global Positioning System (GPS) watch worn by the runner of a four-mile course. Among all the variables that were collected, the relationship between training effect and average heart rate and maximum heart rate is the primary focus. Using heart rate measurements after each run, an analysis of post-exercise heart rate recovery provides an indication of cardiovascular fitness.

ASSIGNMENT TASKS

j. Perform an exploratory analysis to monitor the progress of Kevin.
k. Use visual methods to assess Kevin's progress.
l. Study the pairwise associations between the variables
m. Construct a regression model (or more) to assess the progress of Kevin.

VARIABLES

1 Run: Indicates run number. There were 19 runs in all.
2 Time: How long it took to run the four-mile course, reported in minutes and seconds.
3 Pace: Average time to run one-mile during any one run, reported in minutes and seconds.

4 Calories Burned: Number of calories burned during the four-mile run.

5 Training Effect: Training-induced development of fitness and performance, measured on a scale from 1.0 to 5.0, with categories including Minor (1.0-1.9), Maintaining (2.0-2.9), Improving (3.0-3.9), Highly Improving (4.0-4.9), and Overreaching (5.0).

6 Max HR: Maximum heart rate during the four-mile run, reported in beats per minute.

7 Avg HR: Average heart rate during the four-mile run, reported in beats per minute.

8 Avg Speed: Average speed during four-mile run, reported in miles per hour.

9 Max Speed: Maximum speed during four-mile run, reported in miles per hour.

10 HR Rest: Heart rate immediately after run, reported in beats per minute.

11 HR Rest1: Heart rate one-minute after run, reported in beats per minute.

12 HR Rest2: Heart rate two-minutes after run, reported in beats per minute.

13 HR Change1: Change in heart rate computed as difference in heart rate from start of rest period until one-minute later, reported in beats per minute.

14 HR Change2: Change in heart rate computed as difference in heart rate from start of rest period until two-minutes later, reported in beats per minute.

Note that data values are aligned and are delimited by spaces. There are no missing values.

SUBMITTED BY: Paul J. Laumakis, Department of Mathematics, Rowan University, e-mail: laumakis@rowan.edu.

SOURCE: This data was collected by the runner of a four-mile course using a Garmin Forerunner 610 GPS watch.

Assignments	**24-30**
Title	**Psychometric scales dataset**
Data files	**24-30_duckworth-grit-scale-data (1).zip**
Sample size	**4271**
Number of variables	**~20 covariates - Many items to be used as responses**

This dataset was collected through an on-line personality test. At the end of the personality test, users were asked if their answers were accurate and would be willing to complete an additional survey. At the end of the additional survey users were asked if their answers were accurate and their data could be used for research. This dataset consists of exclusively participants who consented yes at both parts.

This dataset was collected over several pages, the time on each page was recorded:

1. **introelapse**: the time spent on the introduction page to the big five personality test, had an introduction to the big five and a policies statement.
2. **testelapse**: the time spend on the body of the big five personality test
3. **surveyelapse**:the time in seconds spent on the supplemental survey

Technical characteristics:

4. **country** ISO country code
5. **operatingsystem** The operating system of the users computer, determined from HTTP user agent
6. **browser** The browser the user is using, determined from HTTP user agent
7. **screenw** The width of the users screen in pixels, from javascript
8. **screenh** The height of the users screen in pixels, from javascript

Personality test A: big five scales from the international personality item pool.
The following items were rated on a five point scale where 1=Disagree, 3=Neutral, 5=Agree (0=missed).
There are 5 different group of variables measuring different **International Personality Item Pool (IPIP)** personality characteristics (see the attached article for details and file codebook.txt in the attached zip for name descriptions) recorded as

- E1-E10: Extraversion
- N1-N10: Neuroticism
- A1-A10: Agreeableness
- C1-C10: Conscientiousness
- O1-O10: Openness

All were presented on one page in the order E1, N2, A1, C1, O1, E2.....

Then the supplemental survey data were completed (if they agreed to it). Only individuals who opted into the supplemental survey are included in this dataset.
The following items were rated on a five point scale where 1=Very much like me, 2=Mostly like me, 3=Somewhat like me, 4=Not much like me, 5=Not like me at all:
The item text is abbreviated here:
see https://sites.sas.upenn.edu/duckworth/pages/research for the full items.

- GS1-GS12

The following items were presented as a check-list and subjects were instructed "In the grid below, check all the words whose definitions you are sure you know":

- VCL1 boat
- VCL2 incoherent
- VCL3 pallid
- VCL4 robot
- VCL5 audible
- VCL6 cuivocal
- VCL7 paucity
- VCL8 epistemology
- VCL9 florted
- VCL10 decide
- VCL11 pastiche
- VCL12 verdid
- VCL13 abysmal
- VCL14 lucid
- VCL15 betray
- VCL16 funny

A value of 1 is checked, 0 means unchecked. The words at VCL6, VCL9, and VCL12 are not real words and can be used as a validity check.

A bunch more questions were then asked:

- **education**: "How much education have you completed?" 1=Less than high school, 2=High school, 3=University degree, 4=Graduate degree
- **urban**: "What type of area did you live when you were a child?" 1=Rural (country side), 2=Suburban, 3=Urban (town, city)
- **gender**: "What is your gender?" 1=Male, 2=Female, 3=Other
- **engnat**: "Is English your native language?" 1=Yes, 2=No
- **age**: "How many years old are you?"
- **hand**: "What hand do you use to write with?" 1=Right, 2=Left, 3=Both

- **religion**: "What is your religion?" 1=Agnostic, 2=Atheist, 3=Buddhist, 4=Christian (Catholic), 5=Christian (Mormon), 6=Christian (Protestant), 7=Christian (Other), 8=Hindu, 9=Jewish, 10=Muslim, 11=Sikh, 12=Other
- **orientation**: "What is your sexual orientation?" 1=Heterosexual, 2=Bisexual, 3=Homosexual, 4=Asexual, 5=Other
- **race:** "What is your race?" 1=Asian, 2=Arab, 3=Black, 4=Indigenous Australian, Native American or White, 5=Other
- **voted**: "Have you voted in a national election in the past year?" 1=Yes, 2=No
- **married**: "What is your marital status?" 1=Never married, 2=Currently married, 3=Previously married
- **familysize**: "Including you, how many children did your mother have?"

NOTE: for the variable race, an error in the programming of the survey left Indigenous Australian, Native American and White responses all registering as the same value.

References for the the IPIP big five scales (Questionaire A):

- **Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4,* 26-42.**
- http://www.sjdm.org/dmidi/Big_Five_Markers.html
- http://ipip.ori.org/New_IPIP-50-item-scale.htm

References for the Duckworth personality scale (Questionaire B):

- Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087-1101.
- https://upenn.app.box.com/DuckworthPeterson

ASSIGNMENT 24: Construct a single GS response measurement score by considering the sum of items GS1-GS12. Additionally to this response, use all other variables except the IPIP items in your analysis.

ASSIGNMENT 25: Construct a single Extraversion response measurement score by considering the sum of items E1-E10. Additionally to this response, use all other variables except the IPIP items and Duckworth items (GS1-GS12) in your analysis.

ASSIGNMENT 26: Construct a single Neuroticism response measurement score by considering the sum of items N1-N10. Additionally to this response, use all other variables except the IPIP items and Duckworth items (GS1-GS12) in your analysis.

ASSIGNMENT 27: Construct a simple Agreeableness response measurement score by considering the sum of items A1-A10. Additionally to this response, use all other variables except the IPIP items and Duckworth items (GS1-GS12) in your analysis.

ASSIGNMENT 28: Construct a single Conscientiousness response measurement score by considering the sum of items C1-C10. Additionally to this response, use all other variables except the IPIP items and Duckworth items (GS1-GS12) in your analysis.

ASSIGNMENT 29: Construct a single Openness response measurement score by considering the sum of items O1-O10. Additionally to this response, use all other variables except the IPIP items and Duckworth items (GS1-GS12) in your analysis.

ASSIGNMENT 30: Construct a single Word based response measurement score by considering the sum of **the appropriate** items VCL1-VCL16. Additionally to this response, use all other variables except the IPIP items and Duckworth items (GS1-GS12) in your analysis.

FOR ALL ASSIGNMENTS 24-30

a) Explore your data. Describe and visualize the most important characteristics of your dataset.
b) Perform and interpret the appropriate pairwise comparisons.
c) Fit an appropriate model for the response that corresponds to your model. Interpret the results.