

# Package ‘peopleanalyticsdata’

April 13, 2021

**Title** Data Sets for Keith McNulty's Handbook of Regression Modeling in People Analytics

**Version** 0.2.1

**Description** Data sets for statistical inference modeling related to People Analytics. Contains various data sets from the book 'Handbook of Regression Modeling in People Analytics' by Keith McNulty (2020).

**License** MIT + file LICENSE

**Encoding** UTF-8

**LazyData** true

**URL** <http://peopleanalytics-regression-book.org>

**RoxygenNote** 7.1.1

**Depends** R (>= 2.10)

**NeedsCompilation** no

**Author** Keith McNulty [aut, cre] (<<https://orcid.org/0000-0002-2332-1654>>)

**Maintainer** Keith McNulty <[keith.mcnulty@gmail.com](mailto:keith.mcnulty@gmail.com)>

**Repository** CRAN

**Date/Publication** 2021-04-13 10:00:02 UTC

## R topics documented:

charity_donation . . . . .	2
employee_performance . . . . .	3
employee_survey . . . . .	3
graduates . . . . .	4
health_insurance . . . . .	4
job_retention . . . . .	5
learning . . . . .	6
managers . . . . .	7
politics_survey . . . . .	8
promotion . . . . .	9
recruiting . . . . .	9

salespeople . . . . .	10
soccer . . . . .	11
sociological_data . . . . .	11
speed_dating . . . . .	12
ugtests . . . . .	13
<b>Index</b>	<b>14</b>

---

charity_donation	<i>Charity donation data</i>
------------------	------------------------------

---

## Description

Fictional data on the demographics and donation behavior of donors to a wildlife charity

## Usage

charity\_donation

## Format

A dataframe with 354 rows and 8 variables:

**n\_donations** The total number of times the individual donated previous to the month being studied

**total\_donations** The total amount of money donated by the individual previous to the month being studied

**time\_donating** The number of months between the first donation and the month being studied

**recent\_donation** Whether or not the individual donated in the month being studied

**last\_donation** The number of months between the most recent previous donation and the month beng studied

**gender** The gender of the individual

**reside** Whether the person resides in an Urban or Rural Domestic location or Overseas

**age** The age of the individual

## Examples

charity\_donation

---

employee\_performance    *Employee performance data*

---

**Description**

Fictional data on employee performance evaluation metrics for a group of salespeople.

**Usage**

employee\_performance

**Format**

A dataframe with 366 rows and 5 variables:

**sales** The annual sales of the individual in millions of dollars

**new\_customers** The number of new customers acquired by the individual

**region** The region the individuals works in - North, South, East or West

**gender** The gender of the individual

**rating** The performance rating of the individual - 1 = Low, 2 = Middle, 3 = High

**Examples**

employee\_performance

---

employee\_survey    *Employee survey data*

---

**Description**

Fictional data on the results of an engagement survey among company employees on a four-point Likert scale indicating increasingly positive sentiment

**Usage**

employee\_survey

**Format**

A dataframe with 2833 rows and 14 variables:

**Happiness** The employee rating on their overall happiness

**Ben1, Ben2, Ben3** The employee rating on three questions related to employment benefits

**Work1, Work2, Work3** The employee rating on three questions related to general work environment

**Man1, Man2, Man3** The employee rating on three questions related to perceptions of management

**Car1, Car2, Car3, Car4** The employee rating on four questions related to perceptions of career prospects

**Examples**

employee\_survey

---

graduates                      *Graduate salary data*

---

**Description**

Data on graduate salaries in the United States

**Usage**

graduates

**Format**

A dataframe with 173 rows and 5 variables:

**Major** The specific subject major

**Discipline** The broad subject discipline

**Total** The number of graduates of working age in the US

**Unemployment\_rate** The proportion of graduates currently unemployed

**Median\_salary** The current median salary of those employed in US dollars

**Source**

[FiveThirtyEight](#)

**Examples**

graduates

---

health\_insurance                      *Health insurance data*

---

**Description**

Fictional data on the choice of health insurance product by employees of a large company

**Usage**

health\_insurance

**Format**

A dataframe with 1453 rows and 6 variables:

**product** The choice of product of the individual - A, B or C

**age** The age of the individual when they made the choice

**household** The number of people living with the individual in the same household at the time of the choice

**position\_level** The individual's position level in the company at the time they made the choice, where 1 is the lowest and 5 is the highest

**gender** The gender of the individual as stated when they made the choice

**absent** The number of days the individual was absent from work in the year prior to the choice

**Examples**

health\_insurance

---

job_retention	<i>Job retention data</i>
---------------	---------------------------

---

**Description**

Fictional data on the retention of employees in various fields of employment over a 12 month period

**Usage**

job\_retention

**Format**

A dataframe with 3770 rows and 7 variables:

**gender** The gender of the individual studied

**field** The field of employment of the individual at the beginning of the study

**level** The level of the position of the individual in their organization at the beginning of the study - Low, Medium or High

**sentiment** The sentiment score reported by the individual on a scale of 1 to 10 at the beginning of the study, with 1 indicating extremely negative sentiment and 10 indicating extremely positive sentiment

**intention** A score of 1 to 10 reported by the individual at the beginning of the study regarding their intention to leave their job in the next 12 months, where 1 indicates an extremely low intention and 10 indicates an extremely high intention

**left** A binary variable indicating whether or not the individual had left their job as at the last follow-up

**month** The month of the last follow-up

**Examples**

job\_retention

---

learning

*Learning program feedback data*

---

**Description**

Fictional data on feedback from participants in a set of learning programs

**Usage**

learning

**Format**

A dataframe with 4974 rows and 8 variables:

**idcode** The unique ID code of the participant

**rec** A binary value indicating whether the participant would recommend the program to others

**rel** A rating from the participant on the relevance of the program to their work, where 1 is Very Low and 5 is Very High

**fun** A rating on how enjoyable and fun the participant found the program, where 1 is Very Low and 5 is Very High

**clar** A rating from the participant on the clarity of the content and teaching in the program, where 1 is Very Low and 5 is Very High

**home** A rating from the participant on the quality of the homework or project work in the program, where 1 is Very Low and 5 is Very High

**class** A rating from the participant on the quality of the overall class who attended the program, where 1 is Very Low and 5 is Very High

**fac** A rating from the participant on the quality of the program faculty and instructors, where 1 is Very Low and 5 is Very High

**Examples**

learning

---

managers

*Manager performance data*

---

### Description

Fictional data on the performance and other characteristics of a group of managers in a large company

### Usage

managers

### Format

A dataframe with 571 rows and 13 variables:

**employee\_id** The unique ID number for each manager

**performance\_group** The performance group of each manager in a recent performance review: Bottom performer, Middle performer, Top performer

**yrs\_employed** Total length of time employed by the company in years

**manager\_hire** Whether or not the individual was hired directly to be a manager (Y) or promoted to manager (N)

**test\_score** Score on a test given to all managers

**group\_size** The number of employees in the group the manager is responsible for

**concern\_flag** Whether or not the individual has been the subject of a complaint by a member of their group

**mobile\_flag** Whether or not the individual works mobile (Y) or in the office (N)

**customers** The number of customer accounts the manager is responsible for

**high\_hours\_flag** Whether or not the manager has entered unusually high hours into their timesheet in the past year

**transfers** The number of transfer requests coming from the manager's group while they have been a manager

**reduced\_schedule** Whether the manager works part time (Y) or full time (N)

**city** The current office of the manager

### Examples

managers

---

politics\_survey      *Politics survey data*

---

### Description

Fictional data from a survey conducted by a political party on a Likert scale of 1 to 4 indicating increasingly positive sentiment

### Usage

```
politics_survey
```

### Format

A dataframe with 2108 rows and 23 variables:

**Overall** The respondent's overall intention to vote for the party in the next election

**Pol1, Pol2, Pol3** The respondent's sentiment on three questions related to the policies of the party

**Hab1, Hab2, Hab3** The respondent's sentiment on three questions regarding prior voting habits in relation to the party

**Loc1, Loc2, Loc3, Loc4** The respondent's sentiment on four questions related to their interest in local issues

**Env1, Env2** The respondent's sentiment on two questions related to their interest in environment issues

**Int1, Int2** The respondent's sentiment on two questions related to their interest in international issues

**Pers1, Pers2, Pers3** The respondent's sentiment on three questions related to their perceptions of the personalities of local and national party leaders

**Nat1, Nat2, Nat3** The respondent's sentiment on three questions related to their interest in national issues

**Eco1, Eco2** The respondent's sentiment on two questions related to their interest in economic issues

### Examples

```
politics_survey
```



---

promotion

*Promotion data*

---

### Description

Fictional data on promotions in a retail company

### Usage

promotion

### Format

A dataframe with 1134 rows and 5 variables:

**diverse** A binary value indicating membership of a diversity group at the company

**flexible** A binary value indicating whether or not the individual worked part-time for at least 6 months

**store** A binary value indicating whether the individual joined in a position working in the retail stores

**promoted** A binary value indicating whether or not the individual was promoted

**result** The year of the last record of the individual, where the date they joined was year 0. If the individual was promoted, this will be the year of the promotion.

### Examples

promotion

---

recruiting

*Recruiting data*

---

### Description

Fictional data on applicants to a graduate recruiting program in a financial services company

### Usage

recruiting

**Format**

A dataframe with 966 rows and 8 variables:

**gender** The gender of the applicant

**sat** The SAT score of the applicant

**gpa** The GPA of the applicant

**apttest** The result of an aptitude test given to the applicant

**int1, int2** Applicant rating given by two line manager interviewers, on a Likert Scale of 1 to 5 indicating increasing positivity

**int3** Applicant rating given by a human resources interviewer, on a Likert Scale of 1 to 5 indicating increasing positivity

**hired** Binary indicating whether the decision was Hire (1) or No Hire (0)

**Examples**

recruiting

---

salespeople

*Salespeople promotion data*

---

**Description**

Fictional data on promotion and performance for salespeople in a technology company

**Usage**

salespeople

**Format**

A dataframe with 351 rows and 4 variables:

**promoted** A binary value indicating 1 if the individual was promoted and 0 if not

**sales** The sales (in thousands of dollars) attributed to the individual in the period of the promotion

**customer\_rate** The average satisfaction rating from a survey of the individual's customers during the promotion period

**performance** The most recent performance rating prior to promotion from 1 (lowest) to 4 (highest)

**Examples**

salespeople

---

 soccer

*Soccer discipline data*


---

**Description**

Fictional data on disciplinary measures by referees in soccer games

**Usage**

soccer

**Format**

A dataframe with 2291 rows and 7 variables:

**discipline** A record of the maximum discipline taken by the referee against the player in the game. “None” means no discipline was taken, “Yellow” means the player was issued a yellow card (warned), “Red” means the player was issued a red card and ordered off the field of play

**n\_yellow\_25** The total number of yellow cards issued to the player in the previous 25 games they played prior to this game

**n\_red\_25** The total number of red cards issued to the player in the previous 25 games they played prior to this game

**position** The playing position of the player in the game: “D” is defence (including goalkeeper), “M” is midfield and “S” is striker/attacker

**result** The result of the game for the team of the player - “W” is win, “L” is lose, “D” is a draw/tie

**country** The country in which the game took place - England or Germany

**level** The skill level of the competition in which the game took place, with 1 being higher and 2 being lower

**Examples**

soccer

---

 sociological\_data

*Sociological survey data*


---

**Description**

Fictional data on a sociological survey related to income levels in various regions of the world.

**Usage**

sociological\_data

**Format**

A dataframe with 2618 rows and 9 variables:

- annual\_income\_ppp** The annual income of the individual in PPP adjusted US dollars
- average\_wk\_hrs** The average number of hours per week worked by the individual
- education\_months** The total number of months spend by the individual in formal primary, secondary and tertiary education
- region** The region of the world where the individual lives
- job\_type** Whether the individual works in a skilled or unskilled profession
- gender** The gender of the individual
- family\_size** The size of the individual's family of dependents
- work\_distance** The distance between the individual's residence and workplace in kilometers
- languages** The number of languages spoken fluently by the individual

**Examples**

sociological\_data

---

speed\_dating

*Speed dating data*

---

**Description**

Simplified version of the Columbia University speed dating experiment data set

**Usage**

speed\_dating

**Format**

A dataframe with 8378 rows and 11 variables:

- iid** An id number for the individual
- gender** The gender of the individual with 0 as Female and 1 as Male
- match** Indicates if the meeting resulted in a match
- samerace** Indicates if both the individual and the partner were of the same race
- race** The race of the individual, with race coded as follows: Black/African American=1, European/Caucasian-American=2, Latino/Hispanic American=3, Asian/Pacific Islander/Asian-American=4, Native American=5, Other=6
- goal** The reason why the individual is participating in the event, coded as follows: Seemed like a fun night out=1, To meet new people=2, To get a date=3, Looking for a serious relationship=4, To say I did it=5, Other=6

- dec** A binary rating from the individual as to whether they would like to see their partner again (1 is Yes and 0 is No)
- attr** The individual's rating out of ten on the attractiveness of the partner
- intel** The individual's rating out of ten on the intelligence level of the partner
- prob** The individual's rating out of ten on whether they believe the partner will want to see them again
- agediff** The absolute difference in the ages of the individual and the partner

**Source**

Andrew Gelman

**Examples**

speed\_dating

---

ugtests                      *Undergraduate examination data*

---

**Description**

Fictional data on examination scores of undergraduates on a four year biology degree program.

**Usage**

ugtests

**Format**

A dataframe with 975 rows and 4 variables:

**Yr1** Score in the first year examination on a scale of 0-100

**Yr2** Score in the second year examination on a scale of 0-200

**Yr3** Score in the third year examination on a scale of 0-200

**Final** Score in the final year examination on a scale of 0-300

**Examples**

ugtests

# Index

## \* datasets

- charity\_donation, 2
- employee\_performance, 3
- employee\_survey, 3
- graduates, 4
- health\_insurance, 4
- job\_retention, 5
- learning, 6
- managers, 7
- politics\_survey, 8
- promotion, 9
- recruiting, 9
- salespeople, 10
- soccer, 11
- sociological\_data, 11
- speed\_dating, 12
- ugtests, 13

  

- charity\_donation, 2
- employee\_performance, 3
- employee\_survey, 3
- graduates, 4
- health\_insurance, 4
- job\_retention, 5
- learning, 6
- managers, 7
- politics\_survey, 8
- promotion, 9
- recruiting, 9
- salespeople, 10
- soccer, 11
- sociological\_data, 11
- speed\_dating, 12
- ugtests, 13