Topic ideas

Stats is 'Fun': Dav King, Thomas Barker, Luke Thomas, Harry Liu

2022-10-18

```r
library(tidyverse)
# load other packages as needed
```

```r
# load all data sets
players <- read_csv('data/players.csv')
player_values <- read_csv('data/player_valuations.csv')
appearances <- read_csv('data/appearances.csv')
lifeExp <- read_csv("data/Life Expectancy Data.csv")
```

Data Set 1 (soccer)

Introduction and Data

https://www.kaggle.com/datasets/davidcariboo/player-scores

These data sets from Kaggle include soccer player and game data updated regularly from Transfermarkt, a popular online soccer database that uses data scouts to collect data from the match sheets of soccer games. Observations about players include name, age, nationality, height, position, club, preferred foot, and market value, among others. Observations from specific games include each player's contributions such as minutes played, goals, assists, yellow cards, and red cards.

Research questions

How can we use a player's individual characteristics (nationality, age, height, club, etc.) and performance (goals, assists, minutes, red cards, etc.) to predict their market value?

Glimpse of data

```r
# glimpse data set 1
glimpse(players)
```

```
Rows: 25,585
Columns: 21
$ player_id                   <dbl> 254016, 51053, 31451, 44622, 30802, 214776~
$ last_season                 <dbl> 2013, 2013, 2013, 2013, 2013, 2013, 2013, ~
$ current_club_id             <dbl> 855, 23, 23, 3691, 3302, 43, 3302, 3302, 4~
$ name                        <chr> "arthur-delalande", "daniel-davari", "tors~
$ pretty_name                 <chr> "Arthur Delalande", "Daniel Davari", "Tors~
$ country_of_birth            <chr> "France", "Germany", "Germany", "UdSSR", "~
$ country_of_citizenship      <chr> "France", "Iran", "Germany", "Russia", "Sp~
$ date_of_birth               <date> 1992-05-18, 1988-01-06, 1986-01-07, 1981--~
$ position                    <chr> "Midfield", "Goalkeeper", "Attack", "Defen~
$ sub_position                <chr> "Central Midfield", NA, "Centre-Forward", ~
$ foot                        <chr> "Right", "Right", "Right", "Right", "Right~
$ height_in_cm                <dbl> 186, 192, 192, 182, 183, 191, 174, 174, 18~
$ market_value_in_gbp         <dbl> NA, 135000, NA, NA, NA, NA, 270000, 360000~
$ highest_market_value_in_gbp <dbl> 90000, 1130000, 1130000, 720000, 1080000, ~
$ agent_name                  <chr> NA, "NG360", NA, NA, "Pedro Bravo - Consul~
$ image_url                   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
$ url                         <chr> "https://www.transfermarkt.co.uk/arthur-de~
$ club_id                     <dbl> 855, 23, 23, 3691, 3302, 43, 3302, 3302, 4~
$ domestic_competition_id     <chr> "FR1", "L1", "L1", "RU1", "ES1", "SC1", "E~
$ club_name                   <chr> "ea-guingamp", "eintracht-braunschweig", "~
$ club_pretty_name            <chr> "Ea Guingamp", "Eintracht Braunschweig", "~
```

```r
glimpse(player_values)
```

```
Rows: 368,665
Columns: 7
$ player_id                           <dbl> 254016, 254016, 51053, 51053, 5105~
$ date                                <date> 2013-07-22, 2014-01-14, 2008-08-2~
$ market_value                        <dbl> 90000, 90000, 90000, 90000, 225000~
$ datetime                            <date> 2013-07-22, 2014-01-14, 2008-08-2~
$ dateweek                            <date> 2013-07-22, 2014-01-13, 2008-08-2~
$ current_club_id                     <dbl> 855, 855, 23, 23, 23, 23, 23, 23, ~
$ player_club_domestic_competition_id <chr> "FR1", "FR1", "L1", "L1", "L1", "L~
```

```
glimpse(appearances)
```

```
Rows: 1,071,674
Columns: 12
$ player_id          <dbl> 52453, 67064, 67064, 67064, 67064, 67064, 67064, 67~
$ game_id            <dbl> 2483937, 2479929, 2483937, 2484582, 2485965, 248734~
$ appearance_id      <chr> "2483937_52453", "2479929_67064", "2483937_67064", ~
$ competition_id     <chr> "RU1", "RU1", "RU1", "RU1", "RU1", "RU1", "RU1", "R~
$ player_club_id     <dbl> 28095, 28095, 28095, 28095, 28095, 28095, 28095, 28~
$ goals              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ assists            <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ minutes_played     <dbl> 90, 90, 90, 55, 90, 90, 90, 19, 90, 2, 65, 90, 45, ~
$ yellow_cards       <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, ~
$ red_cards          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ date               <date> 2014-08-08, 2014-08-03, 2014-08-08, 2014-08-13, 20~
$ player_pretty_name <chr> "Haris Handzic", "Felicio Brown Forbes", "Felicio B~
```

Data Set 2

Introduction and Data

https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?resource=download

This data set from Kaggle includes a number of variables that will allow us to predict life expectancy. These data were collected by the World Health Organization's Global Health Observatory from 2000-2015. The WHO tracks a number of different variables related to health using their own methodologies, which are optimized for comparability across country and time; thus, while these data may not entirely agree with the data published by individual countries, they are good comparisons with one another. The data include information on the life expectancy during various years in various countries, as well as a number of other potential predictors like adult mortality rate, BMI, infant deaths, alcohol consumption, and more.

Research questions

How much of the variance in life expectancy can be explained by development status, life expectancy, adult mortality, infant deaths, alcohol consumption, GDP & GDP expenditure on healthcare, BMI, education, and population?

Glimpse of data

```
glimpse(lifeExp)
```

```
Rows: 2,938
Columns: 22
$ Country                            <chr> "Afghanistan", "Afghanistan", "Afgha~
$ Year                               <dbl> 2015, 2014, 2013, 2012, 2011, 2010, ~
$ Status                             <chr> "Developing", "Developing", "Develop~
$ `Life expectancy`                  <dbl> 65.0, 59.9, 59.9, 59.5, 59.2, 58.8, ~
$ `Adult Mortality`                  <dbl> 263, 271, 268, 272, 275, 279, 281, 2~
$ `infant deaths`                    <dbl> 62, 64, 66, 69, 71, 74, 77, 80, 82, ~
$ Alcohol                            <dbl> 0.01, 0.01, 0.01, 0.01, 0.01, 0.01, ~
$ `percentage expenditure`           <dbl> 71.279624, 73.523582, 73.219243, 78.~
$ `Hepatitis B`                      <dbl> 65, 62, 64, 67, 68, 66, 63, 64, 63, ~
$ Measles                            <dbl> 1154, 492, 430, 2787, 3013, 1989, 28~
$ BMI                                <dbl> 19.1, 18.6, 18.1, 17.6, 17.2, 16.7, ~
$ `under-five deaths`                <dbl> 83, 86, 89, 93, 97, 102, 106, 110, 1~
$ Polio                              <dbl> 6, 58, 62, 67, 68, 66, 63, 64, 63, 5~
$ `Total expenditure`               <dbl> 8.16, 8.18, 8.13, 8.52, 7.87, 9.20, ~
$ Diphtheria                        <dbl> 65, 62, 64, 67, 68, 66, 63, 64, 63, ~
$ `HIV/AIDS`                        <dbl> 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0~
$ GDP                               <dbl> 584.25921, 612.69651, 631.74498, 669~
$ Population                        <dbl> 33736494, 327582, 31731688, 3696958,~
$ `thinness  1-19 years`            <dbl> 17.2, 17.5, 17.7, 17.9, 18.2, 18.4, ~
$ `thinness 5-9 years`              <dbl> 17.3, 17.5, 17.7, 18.0, 18.2, 18.4, ~
$ `Income composition of resources` <dbl> 0.479, 0.476, 0.470, 0.463, 0.454, 0~
$ Schooling                         <dbl> 10.1, 10.0, 9.9, 9.8, 9.5, 9.2, 8.9,~
```