/\* PROJEKT na zaliczenie \*/

/\* Patryk Wirkus \*/

/\*Tytuł: Wyszukiwanie związków oraz zależności między danymi poszczególnuch zawodników z gry FIFA19.\*/

/\* Źródło zaimportowanej bazy danych: KAGGLE, plik "data" - zawodnicy z gry FIFA 19. \*/

/\* Cel projektu: Głównym celem projektu jest dokonanie analizy porównawczej zawodników z gry FIFA 19,

na podstawie danych zawartych w tabeli, czyli m.in: dane osobiste, umiejętności

piłkarskie, atrybuty fizyczne, a także aktualna wartość rynkowa.\*/

/\*

WSTĘPNA uwaga dotycząca dwóch użytych funkcji.

"Unable to load the Java Virtual Machine. Please see the installation instructions or system administrator."

WYŚWIETLA SIĘ TO PRZY:

- użyciu PROC FREQ do wykresów, tj. 4 użycie tej procedury (ROBIONE NA PODSTAWIE SKRYPTU)

- użcyiu SGPLOT (ROBIONE NA PODSTAWIE HELP)

Po wyjaśnieniu:

Kod tych dwóch procedur jest prawidłowy. Z tym że, prawdopodobnie NIE zainstalowałem pewnej części programu,

stąd te dwie funkcje nie działają. Natomiast, na PG wszystko odziała poprawnie.

\*/

/\* Pierwszym krokiem naszego projektu jest import danych, tu: plik 'data' w formacie csv. \*/

LIBNAME DANE BASE "Z:\Studiav2\SAS\Workspace" ;

/\* LIBNAME DANE BASE "C:\Users\student\Desktop\"; - Tak bym tworzył na PG. \*/

**proc** **import** datafile='C:\Users\Patryk\Desktop\data.csv'

/\* proc import datafile='C:\Users\student\Desktop\data.csv' - Tak bym tworzył na PG. \*/

out=dane.mydata

dbms=csv

replace;

getnames=yes;

**run**;

/\* Pocedura proc contents \*/

**proc** **contents** data=dane.mydata;

**run**;

/\* Do wykonania wykresów będę mógł wykorzystać zmienne dwuczłonowe, wykonując poniższą procedurę. \*/

**data** dane.wykresy;

set dane.mydata;

rename "Weak Foot"n = Weak\_Foot;

rename "Preferred Foot"n = Preferred\_Foot;

rename "International reputation"n = International\_reputation;

rename "Skill moves"n = Skill\_moves;

rename "Jersey Number"n = Jersey\_Number;

**run**;

/\* Wyświetlenie pierwszych 100 obserwacji z zaimportowanej tabeli. \*/

**data** dane.obserwacje;

set dane.mydata(firstobs=**1** obs=**100**);

**run**;

title "100 obserwacji z zaimportowanej tabeli";

**proc** **print** data=dane.obserwacje noobs;

**run**;

/\*

UWAGA, bardzo ważne !!!

Tutaj, za pomocą funkcji NOOBS, wyświetliła mi się w LOGU informacja o dodaniu 100 obserwacji,

czyli w pełni prawidłowo. ALE w wytycznych chodziło o to, by wykonać

"porównanie liczby obserwacji z obserwacjami w pliku źródłowym".

I pojawił się pewien problem, gdyż na początku wykonałem instrukcję:

"proc print data=dane.mydata noobs;

run;"

Wtedy, wyświetlił mi się raport dla tabeli źródłowej, a także informacja w LOGU, że 18207 obserwacji

zostało przeczytanych. Tyle też zawiera tabela źródłowa. Czyli wniosek stąd, że wszytskie wiersze

z importowanego pliku się pobrały. Ale drukowanie tabeli pierwotnej trwało bardzo długo, lecz działało.

Jednak w pewnym momencie wyskakiwał błąd, że tabela wejściowa zawiera błąd, i NIE można wygenerować raportu.

Dlatego też, wykonałem tę samą instrukcję dla mniejszego zbioru. I wtedy wszystko znów zadziałało.

\*/

/\* Użycie 'WHERE', by pokazać umiejętności piłkarzy z krajów, które zwyciężały na dwóch ostatnich

mistrzostwach świata, odpowiednio w roku 2014 i 2018. \*/

**DATA** WHERE;

SET dane.mydata;

WHERE Nationality = 'Germany' | Nationality = 'France';

**RUN**;

/\* Użycie procedury statystycznej - PROC FREQ do:/

/\* 1. Wzrost piłkarzy \*/

title "Rozkład wzrostu wszystkich zawodników";

**PROC** **FREQ** DATA = dane.mydata/\*dane.obserwacje; tak zrobiłem pierwotnie, bo nie chciało działać.\*/;

TABLES height /nocum;

**RUN**;

/\* 2. Zależność potencjału zawodnika od danej narodowości. \*/

title "Zależność potencjału zawodnika od danej narodowości";

**PROC** **FREQ** DATA = dane.mydata;

TABLES Potential\*Nationality;

**RUN**;

/\* 3. Rozkład pensji piłkarzy w stosunku do ich wieku, gdzie format odnosi sie do etapu kariery;

utworzony własny format typu 'value'.\*/

title "Rozkład Pensji piłkarzy w stosunku do ich wieku";

**PROC** **FORMAT** ;

value age

low-**23** = "Poczatek"

**23** < -**32** = "Optymalna forma"

**32** < -high = "Schylek"

;

**RUN**;

title "Rozkład wzrostu piłkarzy w stosunku do ich wieku";

**PROC** **FREQ** DATA = dane.mydata;

TABLES Value\*Age;

format Age age.;

**RUN**;

/\* 4. Wykres przedstawia rozkład wzrostu piłkarzy. \*/

**PROC** **SORT** DATA = dane.mydata;

BY Height;

**RUN**;

ods graphics on;

**PROC** **FREQ** DATA=dane.mydata order=DATA;

TABLES Height /PLOTS(ONLY)=FREQPLOT;

weight Age;

**RUN**;

ods graphics off;

/\* Użycie funkcji matematycznych na wybranych zmiennych, w tym wymagane w wytycznych:

sprawdzenie za pomocą funkcji 'NMISS', czy w zbiorze danych istnieją braki danych na liście wartości

numerycznych.

'Na marginesie': Bez użycia tej funkcji widać, że SĄ braki danych, gdyż przy pierwszym wykonaniu PROC FREQ,

wygenerowała się w raporcie (na samym dole) liczebność braków danych dotycząca wzrostu.\*/

**DATA** funkcje\_matematyczne1;

SET dane.mydata;

**RUN**;

**proc** **transpose** data=funkcje\_matematyczne1 out=funkcje\_matematyczne2;

var Age Overall Height;

**run**;

**data** funkcje\_matematyczne3;

set funkcje\_matematyczne2;

suma=sum(of col1-col18207);

srednia=mean(of col1-col18207);

wartosc\_najmniejsza=min (of col1-col18207);

wartosc\_najwieksza=max (of col1-col18207);

liczba\_brakow=nmiss (of col1-col18207);/\*NMISS-zwraca liczbę braków danych na liście wartości numerycznych\*/

liczba\_NIEbrakow=N (of col1-col18207);/\*N-zwraca liczbę wartości NIEbedących brakami danych na liście

wartości numerycznych\*/

**run**;

/\* WYKRESY \*/

/\* Wykres przedstawia rozkład wieku piłkarzy z automatycznym podziałem lat. \*/

title "Rozkład wieku piłkarzy z automatycznym podziałem lat";

**PROC** **gchart** DATA = dane.mydata;

block Age;

**RUN**;

/\* Wykres przedstawia rozkład wieku piłkarzy ze zmienionym podziałem lat. \*/

title "Rozkład wieku piłkarzy ze zmienionym podziałem lat";

**PROC** **gchart** DATA = dane.mydata;

block Age / midpoints=**16** **18** **20** **22** **24** **26** **28** **30** **32** **34** **36** **38** **40** **42** **44**;

/\*wyszczególniam za pomocą instrukcji 'midpoint' punkty środkowe przedziałów\*/

**RUN**;

/\* Wykres z użyciem opcji i funkcji: Sumvar, Hbar, Pattern, Color. \*/

title "Średni wiek zawodnika względem jego potencjału";

pattern value=empty color=bip; /\*'pattern value=empty'-powoduje rysowanie pustych słupków\*/

/\*color-decyduje o kolorze obramowania słupków\*/

**PROC** **gchart** DATA = dane.mydata;

hbar Potential / sumvar=Age type=mean;

/\*sumvar - definuije zmienną, dla której wartości (średnie/sumy) wyświetla się na wykresie\*/

**RUN**;

/\* Wykres kołowy ukazujący zależność umiejętności gry słabszą nogą piłkarza,

w stosunku do jego preferowanej nogi. \*/

title "Zależność słabszej nogi od preferowanej nogi";

**PROC** **gchart** DATA = dane.wykresy;

pie Preferred\_Foot / sumvar=Weak\_foot type=mean;

**RUN**;

/\* Wykres punktowy z użyciem funkcji gplot. \*/

title "Zależność wieku od pozycji zawodnika";

**PROC** **gplot** DATA = dane.mydata;

plot Age\*Position;

**run**;

/\* Wykres bąbelkowy, tj. z użyciem funkcji 'bubble'. \*/

title "Wiek piłkarza, a jego potencjał i obecne uśrednione umiejętności";

**PROC** **gplot** DATA = dane.wykresy;

bubble overall\*potential=age;

**run**;

/\* Użycie SGPLOT - wykres przedstawiwa ilość zawodników na poszczegolnych pozycjach \*/

**PROC** **SGPLOT** DATA = dane.wykresy;

yaxis label="Ilosc graczy na poszczegolnych pozycjach";

vbar Position;

**RUN**;

/\* Łączenie zbiorów za pomocą 'merge'. \*/

**data** dane.tabela\_łącząca;

merge dane.mydata dane.wykresy;

drop VAR1 ID;

**run**;

/\* Użycie procedury statystycznej - PROC MEANS do:/

/\* 1. Obliczenie 3 wartości (z użyciem f-cji matematycznych), dla zmiennych:

Overall, Potential, Age, które łączy 1 cecha wspólna - klub.\*/

title "Wskazanie wartości statystycznych w poszczególnych klubach danych wartości zmiennych";

**PROC** **MEANS** DATA = dane.mydata

MAX

MIN

MEAN

;

class Club;

var Overall Potential Age;

**RUN**;

/\* 2. Obliczenie 2 wartości (z użyciem f-cji matematycznych), dla zmiennych:

Weak\_Foot Preferred\_Foot International\_Reputation Skill\_Moves, które łączy 1 cecha wspólna - narodowość.\*/

title "Wskazanie wartości statystycznych w poszczególnych krajach danych wartości zmiennych";

**PROC** **MEANS** DATA = dane.wykresy

MIN

MAX

;

class Nationality;

var Weak\_Foot International\_Reputation Skill\_Moves Jersey\_Number;

**RUN**;

/\* GOTO \*/

/\* 1 \*/

**data** goto;

goto = zbior;

set dane.mydata;

zbior: set dane.wykresy;

**run**;

/\* 2 \*/

**data** goto\_1;

label = goto\_1;

goto\_1 = zbior\_1;

set dane.mydata;

zbior: set dane.wykresy;

**run**;

/\* 3 - NAJLEPSZE użycie instrukcji \*/

/\* Celem użycia GOTO w tej tabeli jest zsumowanie wieku wszystkich zawodników z danej nacji. \*/

**data** GermanyGermany(drop=sumaFRA) FranceFrance(drop=sumaGER);

set WHERE;

if Nationality = 'Germany' then goto plik;

sumaFRA+Age;

output FranceFrance;

return;

plik: sumaGER+Age;

output GermanyGermany;

**run**;

/\* LINK \*/

**data** link;

link zbiory;

set dane.mydata;

return;

zbiory: set dane.wykresy;

return;

**run**;

/\* TABLICE \*/

/\* 1 - Rozbudowana próba \*/

**data** TABLICA\_wstęp;

set dane.wykresy(firstobs=**1** obs=**40**);

keep Name Overall Weak\_Foot Skill\_moves International\_reputation;

suma=**0**;

/\*array technika (3) International\_reputation--Weak\_Foot;\*/

array technika (**3**) International\_reputation Skill\_moves Weak\_Foot;

array dobra\_technika (**3**) \_temporary\_ (**3** **3** **3**);

do i=**1** to **3**;

if technika(i)>=dobra\_technika(i) then suma+**1**;

end;

drop i;

**run**;

/\* 2 - Ta na pewno działa \*/

**data** tab\_pomocnicza;

set funkcje\_matematyczne3;

drop suma srednia wartosc\_najmniejsza wartosc\_najwieksza liczba\_brakow liczba\_NIEbrakow;

**run**;

**data** tab\_1;

set funkcje\_matematyczne3;

array kw (\*) \_NAME\_ col1-col18207;

do i=**1** to dim(kw);

kw(i)=kw(i)\***3**;

end;

drop i;

**run**;

/\* MAKRO \*/

/\* 1 - działa na pewno poprawnie \*/

/\* Ma ono za zadanie wyświetlić informacje o CR7 z tabeli 'TABLICA\_wstęp' \*/

%LET pilkarz=Cristiano Ronaldo;

**data** CR7;

set TABLICA\_wstęp;

put "&pilkarz";

where Name = "&pilkarz";

**run**;

/\* !!! Do tego momentu WSZYSTKO działa (tj. NIE wyskakują błędy), więc pozostała część to

(nieudane) dodatki !!! \*/

/\* 2 - robione analogicznie do przykładu pod zadaniem nr\_2 \*/

/\* Ma ono za zadanie wyświetlić raport na temat zawodnika o imieniu i nazwisku danym jako parametr.

Nazwa tego makra to 'pilkarz', zaś parametrem jest 'sportowiec'. \*/

%***JEDEN*** pilkarz(sportowiec);

title "Dane zawodnika: &sportowiec";

**proc** **report** data=TABLICA\_wstęp(WHERE=(Name=&sportowiec));

column Name Club Age Overall;

**run**;

**%mend**;

options mprint;

/\*

A tak (CHYBA) sprawdzam działanie makra:

%JEDEN pilkarz(Neymar JR);

\*/

/\* 3 \*/

/\* Ma ono za zadanie wyświetlić informacje z tabeli 'TABLICA\_wstęp',

ponadto P - numer pierwszej wczytywanej obserwacji, zaś I - ilość obserwacji. \*/

%***DWA*** Napisz(P, I, dane.mydata);

**PROC** **PRINT** data = &dane.mydata (firstobs=&P obs=&I);

**RUN**;

**%mend**;

options mprint;

/\*

" Czego mogę dowiedzieć się analizując ten zbiór danych ? "

Do najważniejszych/najistotniejszych rzeczy możemy zaliczyć:

- wskazanie ilości zawodników reprezentujących nacje, które zdobyły mistzostwo świata w roku 2014 i 2018

- rozkład wzrostu zawodników

- zależność potencjału od narodowości

- pokazanie wpływu wieku piłkarza na jego aktualną pensję

- prezentacja danych statystycznych, takich jak m. in.: średnia oraz wartości skrajne dla:

wieku, uśrednionych umiejętności, a także wzrostu zawodnika

- rozkład wieku zawodników

- zależność potencjału piłkarza od jego wieku

- wyszczególnienie danych statystycznych zawodnika zależnie od reprezenowanego klubu lub narodowość

- zsumowanie wieku wszytskich zawodników reprezentujących nacje, które zdobyły mistzostwo świata w

roku 2014 i 2018

\*/

/\*

" Odpowiedź na pytania dotyczące projektu: "

- Zawodnikiem o najwyższych uśrednionych umiejętnościach są: Lionel Messi oraz Cristiano Ronaldo

- Mamy 2112 zawodników reprezentujący państwa, które zdobyły mistzostwo świata w roku 2014 i 2018

- Najmłodszy zawodnik ma 16 lat, zaś najstarszy 45

- Najniższy 'Overall' to 46, zaś największy to 94

- Największy wzrost zawodnika to 6'9, zaś najmniejszy to 5'1

- Najwięcej mamy piłkarzy w przedziale wiekowym 22-26 lat

- W grze, pozycją na której jest najmniej braków danych jeśli chodzi o każdy kolejny rocznik jest bramkarz

\*/