STTN327: Python project 2024

With all the "hype" of the recent Olympic Games, John Game (a sports analyst) decided to gather information about sportsmen and women taking part in 10 different sports. He wants to determine which factors affect performance in each sport. He realised that it would be quite a job to analyse the data properly since it is quite a large dataset (2000 observations with 33 variables). He decided to use student teams from NWU to assist him in conducting the research.

Each team will conduct the research for **one specific sport**.

John wants each team to delve into their section of the data. An explanation of the data is provided at the end of this document.

Phase 1:

Before building any models, they should perform exploratory data analysis (graphs and descriptive statistics) to ascertain which variables might affect performance in a specific sport. Getting a feel of the shapes of the distributions of the variables, whether there are any obvious outliers, etc, is important.

John doesn't want to see 100 graphs…

He wants to see a **table with a summary** where the team explains the graphs they made and what they saw from their graphs. Where there are interesting results, e.g. a box plot or a scatterplot that indicates a fairly **strong effect** on the performance, the graph can be shown. John also doesn't want to see a large number of the same graphs (e.g., 10 boxplots), but rather a **variety of graphs**.

The careful selection of graphs should be **colourful** with the appropriate titles, labels, etc. Add overlays where necessary. Play around with different interesting graphs.

When writing the code, analysts can use the variable names, e.g. x1, x2, etc. However, when presenting results, e.g. a graph, the sports analyst wants to see the names of variables, e.g. fitness, strength, endurance, etc.

Also, present a **table of descriptive statistics** and interpretations (where appropriate). Indicate the data type of each variable.

Note: John accidentally created two variables that measured exactly the same feature. Ensure that you identify them and delete one of them from the dataset.

Phase 2:

John believes that performance (y) is highly dependent on psychological factors (x23) and gender.

For Y and x23, **create categorical variables** that indicate three levels:

Performance < 50: Low
50 <= Performance < 75 : Medium
Performance >= 75: High

X23 < 50: Low
50 <= x23 < 75 : Medium
X23 >= 75: High

Conduct the appropriate test to ascertain whether **performance is indeed dependent on psychological factors** in the sport you are studying. Support your findings with appropriate **graphs**.

Also, perform appropriate tests to determine if **performance** may be **influenced by gender** (in the sport you are studying).

Support your findings with appropriate **graphs**.

Clearly **state why** you use a particular test / graph and what assumptions you are making. Test them if appropriate.

Phase 3:

Keeping in mind the findings of phase 1 and 2, develop a model Johan can use to predict the performance of an athlete in the sport you are analysing. Ensure that you explain how you build the model. Efficient story-telling is of the essence. John wants to understand how your model evolved. If you decide to drop a variable from a model, explain why. Check assumptions.

Show appropriate outputs, hypothesis tests, graphs, interpretations, etc.

State the final model in mathematical form.

Show how well your model would predict a random person (e.g., what would the predicted performance be of a male soccer player if his psychological factors = 85, fitness = 20, ball sense = 0)? You can select the values of the variables yourself. It is just for explanatory purposes. Remember, you want to convince John that you have a good model!

Phase 4:

Once you have found a fairly good model, look whether you can find anything on the internet that suggests that your significant variables have an effect on performance in the sport you are analysing.

Provide a short summary of evidence in the literature with appropriate references.

Final:

The final presentation will consist of slides of the four phases. A neat Word document containing all the code and output must accompany the final presentation.

Format:

For each phase, you need to create a Powerpoint presentation and come and present it in class.

You have to submit the Python code and its output (in a Word file) as well as the Powerpoint slides on eFundi.

Take note

- You must clearly indicate which team member did what. Everyone must have a significant part that they code, analyse and present.
- Your Powerpoint slides should look professional.
- Make sure that the font size is easily readable.
- Make sure that the colour of the text is readable on the background you chose. (I.e. don't use dark font on dark background or light font on light background.
- A slide should not have full sentences. Use short key words. You will explain the keywords while presenting. Preferably not more than 7 lines per slide, 5 words per line.
- Think how to explain your findings in the best possible manner, using flow diagrams / graphs / tables.
- When you present, make eye contact with the audience.
- Project your voice.
- Don't read from your slides.
- Don't rush through the slides. Ensure the audience understands what you share.

Data:

The data is saved in **sport_perform.csv**

Variable	Label
perform	Performance
type	Sport type
sport	Sport
gender	Gender
x01	Acceleration
x02	Arm strength
x03	Artistry and performance
x04	Balance and centre of mass
x05	Ball sense
x06	BMI
x07	Body alignment and posture
x08	Breathing technique
x09	Decision-making
x10	Emotional regulation
x11	Endurance
x12	Equipment
x13	Fitness
x14	Focus
x15	Height
x16	Lactate threshold
x17	Leg strength
x18	Maximal Oxygen uptake
x19	Mental toughness and focus
x20	Muscle strength
x21	Pattern recognition
x22	Physical fitness
x23	Psychological factors
x24	Speed and velocity
x25	Tactical skills
x26	Takeoff angle
x27	Teamwork
x28	Technical skills

Except for BMI and height, the other variables typically have values between 0 and 100 where a high value indicates that a person has a high level of ability for this particular attribute.

Team members (surnames)	**Sport**
	Ice skating
	Soccer
	Basketball
	Long-distance running
	Long jump
	Tennis
	Volleyball
	Hockey
	Chess
	Swimming