Project Part 2

Due date : Friday 25th March by 11pm. This part is worth 25% of the overall.

You will be docked 5% for each day you are late with submission, in accordance with IT Sligo Marks and Standards. There are three files for this assignment.

1. This pdf : `Projectp2.pdf`
2. The data set Premier League Matches 2019/20: `matches.csv`.
3. Spotify data for Taylor Swift's songs: `TaylorSwiftTracks.csv`.

Ensure when you are uploading your work you only upload your own work and **not** any dataset. **Just upload the ipynb file, renamed to have your name and Student Number**.

This is **NOT** a group assignment. You must complete the assignment **individually**. If you share your work, your marks will be divided. Some students may be asked to demonstrate their knowledge of the submitted work. Ensure you read the complete document.

When you are asked to answer a question. Put answers in markdown cells. The interpretation of information and results is at least as important as the building of models.

Hypothesis Tests

Using the matches dataset (that you previously worked on) and Taylor Swift lyrics dataset, you want to use appropriate Hypothesis tests to answer some questions about the dataset.

Do the full steps of each Hypothesis test, explaining in markdown cells

1 Wins/Losses Pre-Post Covid (7 marks) - Football Dataset

The season was broken up due to Covid-19 with the season being paused in March without all the games being completed. Plans were made to restart the season without fans in the stadia. There were calls for the season to be ruled "Null and Void" due to this, with some teams saying they will have lost the home advantage having the play games without fans.

Is the amount of Home Wins/Away Wins/Draws affected by the teams not playing in front of fans?

Complete a full Hypothesis test (as described in notes), using an appropriate method to analyse the set.

Here are some tips to complete this Hypothesis test

1. You are going to want to read in the CSV file with the data being parsed correctly. Use `pd.read_csv?` to see the options. Maybe `parse_dates` and/or `dayfirst` might come in handy.
2. This is so that you can split the data frame in two, pre and post covid. Have these in separate dataframes.
3. Now you need to do a count of Home Wins, Away Wins and Draws both pre and post covid. `FTR` column may be useful for this.
4. With these counts you could now be able to do a Hypothesis Test

2 Difference between Rerecording of Songs (7 marks) - Taylor Swift Dataset

Due to a dispute about the owning of her original masters recordings, Taylor decided to rerecord 6 of her albums so that she would regain control of the masters. So far two of those albums have been re-recorded.

When recording music, changes may happen due to the studio, the singer's voice, or any number of factors that slightly alter the end result of the songs.

Your job, is to pick one of the two albums that have been rerecorded (Fearless or Red) which are in the dataset. The ones marked as Taylor's Version are the new ones. You can only analyse the songs that the original and the remake have in common. Note in the workbook which album you are doing.

Complete a full Hypothesis test on the data to determine if there is enough statistical evidence that the versions do differ. You may have to make two or more hypothesis tests to gather enough evidence.

Write a small paragraph detailing the conclusion you make and the choices you made in coming to that conclusion.

Regression

Now you want to build a couple of Linear Regression models to make predictions

3 Correlation/Regression in songs (6 Marks)

Take the Taylor Swift Spotify dataframe again. You want to investigate if there is any correlation between columns and then build a simple linear regression model to make predictions.

Tips:

1. Take a subset of the columns, getting rid of things that won't have correlation such as dates, names etc.
2. Use pandas to search for correlations and plot a scattermatrix. Use this to choose which columns you are going to focus on.
3. Pick which is the dependent and independent variable, separate them out and build a linear regression model using that data
    - Use whatever library you want
4. Get the Mean-Squared Error and R^2 of the model.
5. Use the model to make some predictions
6. Plot the model line vs some actual data.
7. Comment on how well the model performs.

4 Regression in Diabetes (5 Marks)

Now you are going to build a Multiple Linear Regression Model using a diabetes dataset built into Python.

```
from sklearn import datasets
dataset = datasets.load_diabetes(as_frame=True)
```

1. Write a little bit about what is contained in that dataset and what you are trying to predict
2. Only use columns 0,1,3,4,5 to build your model
3. Fit your model.
4. Get the R^2 score.
5. Comment on how well you think it will perform.
6. Make some predictions
7. Write a summary about the whole thing.

Suspected Cheating & Plagiarism

Attention is drawn to IT Sligo's definition of plagiarism at `https://vle.itsligo.ie/course/view.php?id=2206`. Student plagiarism occurs when a student presents the work of another as their own work, without appropriate acknowledgement. It can include:

- Presenting work which has been copied from the Internet, books, journals or other sources;
- Presenting work which paraphrases the writings of other authors, without acknowledgement;
- Presenting work which has been written by somebody else, such as another student or a family member;
- Presenting work which has been purchased from an Internet site or other source and submitting as own work;
- Presenting work which has been produced collaboratively as one's own individual work;
- Student plagiarism can also occur where a student submits the same piece of their own work for a number of different assignments;

Plagiarism by students can be deliberate, where a student intentionally attempts to pass off the work of another as their own work. It can also be accidental, where a student fails to use appropriate citation and referencing in their work or is unaware of what constitutes plagiarism.