R-COURSE PROJECT

The aim of the course project is demonstrate proficiency in the techniques covered in this class and apply them to a novel dataset in a meaningful way.

To achieve this, students would have to:

- **import** a real life **data set**
- **assess, clean and tidy the data,** and
- perform basic **exploratory data analysis**;
- all while using **R Markdown** to produce an **HTML report** that is fully reproducible.

Students report are expected to tell a story with the data and provide a coherent explanation of their findings. The project is open ended and there is no limit on what tools or packages to use.

PROJECT DATA

The final project for this course will consist of the statistical analysis of any of the provided or downloaded datasets.

All data sets contain key attributes that will demonstrate the capabilities learned in this course. Students will also learn new skills not taught to accomplish this projects.

These include working with:

- Multiple data types (numerics, characters, dates, factors)
- Unclean data (missing values)
- Variables that need to be created.
- Data that needs to be filtered out
- and much more!

You can choose from one of the following data sets:

1. Data on the population of forcibly displaced persons
2. Data on spam emails
3. Data on every world cup match from 1930 to 2018
4. Data on netflix shows
5. Any manageable dataset from TidyTuesday

PROJECT REPORT

1. **Students will write an R Markdown HTML report based on the sections below:**
   A. Introduction:
      - Provide an introduction that explains the problem statement you are addressing.
   B. Packages Required:
      - All packages used are loaded upfront so the reader knows which are required to replicate the analysis.
      - Explanation is provided regarding the purpose of each package.
2. **Data Preparation:**
   - Original source where the data was obtained is cited and hyperlinked.
   - Source data is thoroughly explained.
   - Data importing and cleaning steps are explained in the text and follow a logical process.
   - Once your data is clean, show what the final data set looks like.
   - Provide summary information about the variables of concern in your cleaned data set.
3. **Exploratory Data Analysis**
   - Uncover new information in the data that is not self-evident (slice and dice the data in different ways, create new variables for more information).
   - Provide findings in the form of plots.
   - Plots are carefully presented. One graph illustrates one primary point and is appropriately formatted (plot and axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).
   - Insights obtained from the analysis are explained.
4. **Summary**
   - Summarize the problem statement you addressed.
   - Summarize how you addressed this problem statement (the data used and the methodology employed).
   - Summarize the interesting insights that your analysis provided.
5. **Formatting**
   - Tools and techniques from the course are applied competently.
   - Rmd fully executes without any errors and matches the HTML report submitted by student.

Final project should include:

1. RMarkdown file (formated to clearly present all of your code and results)
2. HTML file
3. Dataset