Welcome! Thanks for taking the time to work on this data science assignment from BetterUp. Below, you'll find the assignment guidelines and instructions. Good luck! We're really looking forward to seeing what you come up with.

Guidelines

- Choose either Assignment A or Assignment B (not both).
- You have 48 hours from the time you download this document to complete your chosen assignment.
- You may use as much of the full 48-hour period as you like to complete the assignment, but we're only expecting a level of quality roughly commensurate with 5-6 hours of work.
- Please return a single archived file (.zip or .tar.gz) containing all of your work, or links to your work, to andrew.reece@betterup.co, before the close of the assignment period.
- Tasks marked with ** are considered challenging. Prioritize getting the basic tasks done first, then go for the challenge items with time remaining.
- It's fine if you don't get all your answers accepted/upvoted (for Assignment A), or if you don't finish all of the tasks (for Assignment B). Just do what you can, and make sure you showcase your skills as a data scientist. If you complete the entire assignment, great!
- Please don't share the details of these assignments with anyone outside of BetterUp. That includes the datasets provided for Assignment B.

Instructions

Assignment A: Answer three questions on StackOverflow.

- Answers posted prior to this assignment date don't count.
- Each of your answers should either be accepted or receive at least one upvote.
    - Exception: If you find an older post that isn't getting much attention, but which you feel would be a good showcase of your talents, feel free to include that as one of your answers (but not more than one). An "older post" means a question originally posted more than a month ago.
- Questions must be tagged with at least one of the following: `R` or `Python` (or variants including but not limited to: `tidyverse`, `dplyr`, `pandas`, `numpy`, `sklearn`).
- You can answer any questions you like, but keep in mind that we'll be using your answers to learn about your understanding of data science problems, and your ability to communicate solutions effectively. (E.g., If you see someone just forgot to add a comma, you might consider skipping that one for the purposes of this assignment.)

- If your answers are extremely short, or heavy on code and light on explanation, consider including a 1-2 page write-up, explaining the problem, how you went about crafting a solution, and why you chose to include these answers for your assignment.
- ** Collect a total of 5 upvotes for your answers (you can answer more than 3 questions to accomplish this).
- ** Answer at least one question with an R tag and at least one question with a Python tag.

Assignment B: Analyze a Twitter dataset.

- Complete this assignment using R or Python.
- Your marketing team wants to understand how Twitter users' behavior changes, based on user age. They've asked you to look at some tweet data and tell some simple descriptive stories, as well as see if you can come up with a way of predicting user age based on other characteristics.
- You have the following datasets to work with:
    - *ages_train.csv*: The training data, indicating the known age for each user ID in the set.
    - *ages_test.csv*: The test data (a set of Twitter user IDs for which the user's age is not known). The goal of the prediction task is to provide as accurate as possible a prediction of the ages of each user in this set.
    - *age_profiles.json*: Twitter user profiles corresponding to the users in the training and test sets. The data format is documented here.
    - *age_tweets.json*: Recent tweets from the users in the training and test sets. The data format is documented here.
    - *mentions.csv*: A data set indicating users that have recently been mentioned by the users in the training/test set in tweets.
    - *mention_profiles.json*: Twitter user profiles corresponding to the users mentioned in *mentions.csv*. The data format is documented here.
    - *friends.csv*: A data set indicating users that the users in the training/test set are following.
    - *friend_profiles.json*: Twitter user profiles corresponding to the users in *friends.csv*. The data format is documented here.
- **Descriptive tasks**
    - Make histograms of followers count, friends count, favorite count, and status count, all of which are in *age_profiles.csv*.
        - Are friend and follower counts correlated?
        - Are favorite and status counts correlated?
        - What stories can you (speculatively) tell about the relationships between any of these variables, based on your findings?
    - Which time zone has the highest proportion of known iOS users in *age_profiles.csv*? Which time zone has the highest proportion of Android users?
    - Use the "mentions" data in *mentions.csv* to come up with a list of Twitter handles that were mentioned by more than one user.

        - Build a list of the top 20 handles (rank by greatest number of unique users mentioning a given handle).
        - Which actor/actress in this top 20 list starred in the Harry Potter movies, and how many unique users mentioned this star's Twitter handle?
    - Break down the sample by age-decade (age 10-20, 20-30, etc).
        - Make a bar chart of age group sample size (x-axis: age group, y-axis: per-group sample size)
        - **Which age group uses the most emojis in their profile status?
        - **Which is the most common emoji?
    - Submit your findings as a PDF. For this section, a formal written report isn't necessary, just provide your answers to each question, along with any associated graphs.
- **Prediction tasks**
    - Using the training data in *ages_train.csv*, build a statistical model to predict the age of users in *ages_test.csv*, based on any of the information provided (profiles, friend networks, and mentions).
        - You may approach this task with age formulated either as a continuous or categorical/ordinal variable. Justify your choice in your report.
        - If you formulate age as a categorical/ordinal variable, create at least five category levels (e.g. age ranges) in your response variable.
    - **Include some measure of user tweet sentiment as a predictor.
        - It's up to you how to measure sentiment, and what kind of feature to build from this measurement. Justify your choices in your report.
    - **Include some measure of emoji use as a predictor.
        - It's up to you what kind of feature to build from this measurement. Justify your choice in your report.
- **Submission guidelines**
    - Submit your code.
        - If you use R, submit your code in `.R` or R Markdown (`.Rmd`) format.
        - If you choose Python, submit your code in `.py` or Jupyter notebook (`.ipynb`) format.
        - You may choose to use whichever version R or Python you feel most comfortable with.
    - Submit your predictions as a separate file, *ages_pred.csv*.
        - This is a comma-separated file with two columns, `ID` and `Age`.
        - The `Age` column will contain predicted ages for each Twitter user ID in the `ID` column. (This is basically the same file as *ages_test.csv*, but with an `Age` column added.)
    - Write up your findings in a 1-2 page report. Graphs can take an additional 1 page if needed.
        - Submit your report as a PDF.
        - The audience for this report is the executive director of a data science team; she will report the results to your Marketing team. That means you

should assume audience familiarity with the technical methods you've used, but focus on telling a clear story. Use minimal reliance on technical terminology to get your point across, as eventually your work will be consumed by a non-technical audience.

- It's up to you to decide how to write up the report, but you should include answers to most or all of these questions:
    - Why did you choose the modeling strategy you chose?
    - How well were you able to predict user age?
    - What were the most important predictors?
    - What, if any, challenges did you encounter with data quality or data manipulation?
    - What, if any, are key limitations to your findings?
    - Would you feel confident in telling your company's marketing team that you can accurately target specific user ages with your model? Why or why not?
    - What's the key takeaway from this project that you'd give to a non-technical audience?