Comprehensive Project

Using all five standard rules that applied to Project 13, in a single R Studio project, do the following:

- For each of the tasks listed below, use data from the *Professional Worker Career Experience Survey, United States, 2003-2004* from http://www.icpsr.umich.edu. You will need to be on campus.
- For each of the tasks listed below, do each of the following:
    a. Conduct the requested analysis, adhering to any data cleaning and assumption checks. This includes transforming any skewed variables as appropriate.
    b. Report all relevant statistical significance tests and effect sizes, regardless of significance.
    c. Include comments (using ##) throughout explaining what you are doing, why you are doing it, and what you interpreted/concluded from what you did.

File 1 – psych.R

- Conduct an exploratory factor analysis of the 9 job satisfaction facets measured (i.e., satisfaction with salary through satisfaction with work itself) to extract 3 factors. Use the *psych* package for this. Identify items belonging to a general factor of job satisfaction based upon your judgment and use those as the variable of interest for the next two analyses (which here will be called: general job satisfaction).
- Test if a person's career field classification predicts general job satisfaction using an ANOVA framework. Also export an APA-style ANOVA summary table to a file called **anova.doc**. Visualize with a horizontal bar chart.
- Report and interpret coefficient alpha for all items measuring intention to leave for another employer, also using the *psych* package. Using the mean score of these items, test if a person's overall stress management score and general job satisfaction interact to predict their intention to leave for another employer using an OLS regression framework. Also export an APA-style regression summary table to a file called **regression.doc**. Visualize by splitting the dataset into above-and-below median scores on job satisfaction, and create a scatterplot with regression lines.

File 2 – datascience.R

- Compare the people in your dataset who have the word "manager" (any capitalization) in their typed-out job titles to people that don't in terms of their overall work family conflict. Visualize with a bar chart.
- Test if the words in a person's job title can be used to predict their overall work-family conflict score using a machine learning framework. Compare at least two machine learning algorithms to address this question. Visualize the comparison. Remember to simplify your data.

File 3 – advisor.Rmd

- Create a PDF stepping your advisor through the analyses you conducted in **psych.R**. Be sure to include comments explaining each step, describing what you concluded from it, and why. Be sure to move headings and comments outside of R chunks into the main PDF.

File 4 – app.Rmd

- Create an app that allows users to select among career field classifications and view a histogram of general job satisfaction within that classification. Also display the mean and SD of the selected group. Paste a comment with a link to the app on shinyapps.io at the end of **datascience.R**.

Supplemental Details on EFAs

- The basic question answered by an exploratory factor analysis is this: if we assume there are a limited number of latent traits or characteristics that created these data, how many traits could have reasonably caused the observed scores on this set of variables?

- For example, let's say five people complete a survey with five questions ranging from 1 to 10:

  Person 1: 4 4 9 9 9
  Person 2: 1 1 3 3 3
  Person 3: 9 9 6 6 6
  Person 4: 3 3 1 1 1
  Person 5: 5 5 6 6 6

  You would probably conclude "oh, items 1 and 2 seem to be caused by the same thing, and items 3, 4 and 5 seem to be caused by another thing." EFA is a mathematical process by which to make that conclusion, rather than relying upon your intuition.

- When you run an EFA, one of the pieces of output you get is called a **pattern matrix**. The numbers in a pattern matrix represent the unique contribution of each variable to each identified factor. For example, with the data above, you might get back a pattern matrix like this:

	Fac1	Fac2
Item 1	.80	.18
Item 2	.70	.13
Item 3	.15	.95
Item 4	.20	.85
Item 5	.19	.75

  These numbers can be interpreted as the regression coefficient if you were to regress Fac1 on Item1-Item5, thus:

  $y'(Fac1) = .80*I1 + .7*I2 + .15*I3 + .20*I4 + .19*I5$
  $y'(Fac2) = .18*I1 + .13*I2 + .95*I3 + .85*I4 + .75*I5$

  From this, you can interpret that Factor 1 is made up primarily of Items 1 and 2, whereas Factor 2 is made up primarily of Items 3, 4 and 5. You could then create scale scores by calculating the means of each set of items. The interpretation of those factors is ultimately up to you as the researcher.