ISM 6361 - Data Mining Analytics & Visualization

Kait Hemphill & Brandon Purvis

Data Viz Project 2 – R

Write Up

In our canvas submission, you will find the following files besides this writeup.

1. Hemphill\_Purvis\_**Poster**\_LaBrieProject2.Rmd
   1. This .Rmd file will knit to an html poster which allows for better zooming and pixilation.
   2. Knits to a .html file
   3. Other necessary files:
      1. Horizontal.png - Necessary image for the tableau comparison in poster.
2. Hemphill\_Purvis\_**PosterPdf**\_LaBrieProject2.**pdf**
   1. Pdf version of the poster, per the requirements. Html is the indented functionality, but Pdf will get the main points across.
3. Hemphill\_Purvis\_**Analyses**\_LaBrieProject2.Rmd
   1. This is a file containing the regression output / calculations we explored.
   2. Knits to a .doc file
4. Hemphill\_Purvis\_**Dashboard**\_LaBrieProject2.Rmd
   1. Dashboard also knits to a html – which will be posted to GitHub (link below)
   2. Knits to a .html file
      1. Other files affiliated:
         1. newdf.csv
         2. dfgrouped.csv
5. Hemphill\_Purvis\_**Animation**\_LaBrieProject2.Rmd
   1. Animation for a scatterplot showcasing the relationship between Resilience and Stress by Age Categories.
   2. Knits to a .html file
6. AIAscored\_dx.csv
   1. Dataframe scored through Available item analysis (AIA; Parent, 2013). Once uploaded, typically renamed as “df”, or otherwise.

GitHub Link for Dashboard:

HTML Link for Poster:

HTML Link for R Markdown Output:

1. State a business reason for selecting your tools (problem you would like to solve).

COVID-19 stay at home orders have offered powerful opportunities to study resilience. The highly stressful conditions brought by the pandemic allows us to consider the extent to which people are adopting new habits. Resilience is defined as not merely surviving adversity, but bouncing back, growing from it, and increasing one’s capacity to face future challenges (Bonanno & Diminich, 2013; Smith et al, 2008).The resilience practices considered Time Management (Carver, 1997), Growth Mindset, (Carver, 1997; Dweck, 2006) and Sleep (Marinus et al., 2003). We will consider these practices alongside perceived stress (Cohen et al., 1983).

1. Document how/where you got your data (if it is publicly available, or internal for a work project).

Participants were reached via Prolific and received $2.00 compensation upon completion of a 15-20-minute survey. Participants (N = 500) were required to be at least 20 years old. The sample was 54.6% male and 44.6% female, and 1.6% other, ranging from 20 to 76 years old (M = 36.22, SD = 12.50). Data was cleaned in R.

Resilience: 6-item scale called the Resilience Assessment Brief (Smith et al., 2008). Coping Skills: Growth Mindset & Time Management (Kearns & Gardiner, 2007).): Brief COPE scale developed by Carver (1997) with adaptations (Dweck, 2006) Stress: 5-item questionnaire (Cohen et al., 1983). Sleep: 2 item scale measuring quality and habits, adapted from the SCOPA (Marinus et al., 2003).

*Data Cleaning and Missingness Dx*

Available item analysis (AIA; Parent, 2013) is a strategy for managing missing data that uses available data for analysis and excludes cases with missing data points only for analyses in which the data points would be directly involved. Parent (2013) suggested that AIA is equivalent to more complex methods (e.g., multiple imputation) across several variations of sample size, magnitude of associations among items, and degree of missingness. Thus, we utilized Parent’s recommendations to guide our approach to managing missing data. Missing data analyses were conducted with the R packages mice (v. 3.7.0), Amelia (v. 1.7.6), and BaylorEdPsych (v. 0.5). We began by deleting cases where missingness was 90% or more. Of the cases remaining, missing values represented .01% of the cases; 99% of the cases had non-missing data. When running the mice package on the remaining cases, there was no need to, as the mice package indicated all were observed.

Scales were calculated using Parent’s recommendation that some reasonable amount of missingness be allowed. Thus, for scales containing only two items, we allowed up to 50% missingness; for scales containing four items, we allowed up to 25% missingness; and for all others, we permitted up to 20% missingness. Applied at this measurement-model level of analysis, Little’s MCAR test, which diagnoses whether or not the missing observations are missing completely at random suggested that our larger p-value (> 0.05) indicated weak evidence against the null hypothesis, so we will fail to reject the null hypothesis, in this case the null hypothesis is that the data is MCAR, no patterns exists in the missing data ( χ2 (22) = 11.98, *p* = .959).

\*All analyses were completed in R Studio (v. 1.2.5033) with R (v. 3.6.2).

1. **Document how you used your tool.  Many tools are super rich in features and you probably will not be exploring all the features, explain the parts you did use.**

Please see R Markdown for a thorough documentation of how we used our tool.

*\*\*The following section will answer both questions 4+5 in the same text.*

1. **Explain why you chose which visualization/charts**
2. **Give an explanation/analysis of the output. What did you learn or uncover?**

Latex Equation Vis – Regression

* Using R to represent the mathematical equation is visually something I appreciate. Sometimes mathematical notation can look rough, clunky, and downright ugly. The way the R translates this visually is beautiful, and I believe it is important to let the reader know that we used regressions to consider the relationships between our variables.

Figure 1– Scatterplot: Stress and Resilience Color was used to indicate the

* This scatterplot with a best fitted line was chosen to plainly show the relationship between the two variables. Consider what this is telling us – that the more stress you have, that is accompanied by lower resilience. For those of us who believe we are invincible or are unaffected by stress, may want to take a closer look at how powerful stress is.
* When stress was regressed on resilience, we see the following:

Coefficients:

Estimate Std.Error t value Pr(>|t|)

(Intercept) 5.06377 0.11060 45.79 < .000 \*\*\*

Stress -0.66119 0.03816 -17.33 < .000 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7235 on 487 degrees of freedom

Multiple R-squared: 0.3814, Adjusted R-squared: 0.3801

F-statistic: 300.2 on 1 and 487 DF, p-value: < 2.2e-16

* When stress is set to zero, Resilience is at pretty high levels *(B* = 5.06; Resilience was on a scale from 0-5). Now, throw stress into the mix and for every 1 unit increase on stress, Resilience goes down by a -0.66. This is a large relationship.
* This means that, when life is good – no stress, no pandemic, no civil unrest, people have decently high resilience in general. We are not in that state – so what can we do?
* We chose this viz/chart because I wanted to showcase the obvious and simple linear relationship, and to showcase that regression is not scary (while the equation might seem so, it is not – it is a line, a perfectly interpretable line). We have considered this graph and code a calculation.

Figure 2 – Scatterplot: Sleep and Stress

* Another simple regression plot here between the relationship of Seep and Stress. We were curious about exploring the relationship between Stress and sleep because we have much reason to believe that stress impacts our sleep.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.76881 0.03530 78.43 <2e-16 \*\*\*

Sleep -0.38413 0.03788 -10.14 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7807 on 487 degrees of freedom

Multiple R-squared: 0.1743, Adjusted R-squared: 0.1726

F-statistic: 102.8 on 1 and 487 DF, p-value: < 2.2e-16

* When sleep is set to zero, Stress is at moderate levels *(B* = 2.77; Stress was on a scale from 0-5, low stress to high stress). Now, factor sleep into the equation - for every 1 unit increase on sleep, Stress goes down by a -0.38.
* Refer to the justification in Figure 1’s section as to why we chose a scatterplot with a best fitted line. We have considered this graph and code a calculation.

Figure 3 – ﻿Scatterplot: Sleep and Resilience

* Another simple regression plot here between the relationship of Seep and Resilience. We were curious about exploring the relationship between Resilience and sleep because we have considerable agency around sleeping and it is suggested to be a highly important resilience practice.

Coefficients:

Estimate Std.Error t value Pr(>|t|)

(Intercept) 3.23302 0.03925 82.369 < .000 \*\*\*

Sleep 0.32625 0.04212 7.746 < .000 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.868 on 487 degrees of freedom

Multiple R-squared: 0.1097, Adjusted R-squared: 0.1079

F-statistic: 60 on 1 and 487 DF, p-value: 5.559e-14

* When sleep is set to zero (terrible sleep in both duration and quality), Resilience is at a moderate level *(B* = 3.23; Resilience was on a scale from 0-5). Now, throw Sleep into the mix and for every 1 unit increase on sleep, Resilience goes up by *B =* .33. This is a positive and linear relationship, also statistically significant.
* Refer to the justification in Figure 1’s section as to why we chose a scatterplot with a best fitted line. We have considered this graph and code a calculation.

Figure 4 – ﻿Correlation Map

* So what about adding a few other variables into the mix beyond sleep? There must be other resilience practices beyond sleep… right? Yes, there are. Specifically, we want to see if age (categories) has anything to do with resilience, stress and sleep, but also – we want to see if two Coping skills such as Time Management and having a Growth Mindset have an impact on any of the above relationships.
* Color here is used to indicate both direction of relationship (positive or negative) as well as indicate strength (magnitude).
* Stress is inversely correlated with age (red), yet age is positively correlated with everything else. This suggests that age is changing the relationship between stress and the other variables. INTERESTING. Our next step will be to put Age into categories and see if we can identify any
* We chose this vis/map because it is a great visual representation of the correlations between all our variables.

Figure 5 – Model of The Impact of Stress on Resilience by Age Group

* To see the relationship stress has on different age categories, we placed age into three different categories: Young Adult (20-34), Adult (35-49), Senior (50-76). We know that these are not typically the “ages of senior,” however, splitting it larger would have created a large difference in subgroup n size and we wanted to avoid that. In general, older people are at higher risk when considering COVID-19, and we thought it would be interesting to probe the relationships between age and stress as it relates to resilience.
* We chose this visual because it is side by side and easy to see that in general, when you have more stress in your life, you’re less resilient. That said, we see what might be considered a curvilinear relationship at the end of both Adult and Senior where the end of their lines start to go up again. If we wanted to explore this relationship further, we would use the following equation. That said, data points are limited for high stress, which increases error in our predictions at the high level.
* Lastly, we notice here that as age goes up, the amount of stress one has is going down. Age is acting similarly to all other variables, which tells us we might want to explore it a little further.

Figure 6 – ﻿Violin Plot 1 - Resilience by Young Adult, Adult, and Senior

* The violin plot is unique to R in that it is not in Tableau, we will compare this in Figure 8.
* Like the Boxplot, a violin plot is a method of plotting numeric data. With the addition of a density plot on each side that shows the probability density of the data at different values (smoothed).
* We chose this viz/chart because it is unique and we wanted to see density as well as mean scores by age category. This combined those aspects uniquely, and it looks aesthetically pleasing.

Figure 7 – ﻿Violin Plot 2 – Sleep by Age category

* Taking the violin one step further, we wanted to see if the relationships and mean values we saw in our first violin plot held true when we considered sleep instead of stress. This visualization is a little unique to interpret because of the z-scoring of the sleep variable (see lessons learned), however, we see that because we standardized, the mean is set to zero. The axis on the left then goes from -3 to +3. What we are seeing here is there is more variability in the Adult + Senior groups to cause bigger differences. We see that the Young Adult group hugs the mean a little more tightly. We also see that there is a larger density of young adults at the half standard deviation above sleep.
* With this one we went with black and white and no gridlines to change it up. The age categories are the same, and standardizing does not change the nature of the data, simply a linear transformation of the scaling.

Figure 8 – ﻿Tableau Comparison – Boxplot

* Similar to the boxplot, we felt that violin plots had an advantage here. Unlike boxpots with means and error bars, violin plots contain all data points. This make them an excellent tool to visualize samples of small sizes (our subgroup n sizes varied). Thus, Violin plots are great when looking at non-parametric numbers, and appropriate even if your data do not conform to normal distribution.
* Reading the violin shape is exactly how you read a density plot: the thicker part means the values in that section of the violin has higher frequency, and the thinner part implies lower frequency. This is something you do not get with a box plot, in fact, box plots seem a little one-dimensional after working with a violin plot.
* One might ask, why don’t we just use a density plot instead of violin plot? When there are too many groups (especially more than 3), their overlapping density plots become difficult to read. This doesn’t happen with violin plots, because violin plots are visually intuitive and attractive to read because they, unlike box pots, do not sit on top of one another and create confusion about what colors we are looking for.

6. Discussion / Application / Lessons Learned

**Discussion**

* **Stress** heavily impacts our ability to maintain resilience – roughly 38% of the variance (Figure 1 output)
* Stress and Resilience differ by Age. Young adults experience more Stress than adults and seniors.
* Growth Mindset and Time Management differ by age as Adults and Seniors more readily adopt a growth mindset, whereas young adults tend not to.
* Having a Growth Mindset under stress, across all ages, predicts resilience.
* Having all these variables in the model soaks up about 44% of the variance in Resilience, which is pretty unreal. That is a large multiple R2

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.743226 0.221812 16.876 < 2e-16 \*\*\*

Stress -0.567423 0.041336 -13.727 < 2e-16 \*\*\*

Sleep 0.071819 0.037120 1.935 0.0536 .

Age 0.006708 0.002597 2.583 0.0101 \*

CopeTime 0.040472 0.060183 0.672 0.5016

CopeGrowth 0.215751 0.049932 4.321 1.89e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6884 on 479 degrees of freedom

(4 observations deleted due to missingness)

Multiple R-squared: 0.4419, Adjusted R-squared: 0.4361

F-statistic: 75.86 on 5 and 479 DF, p-value: < 2.2e-16

**Application**

* One might consider reframing during negative circumstances - trying to look for something good in what is happening.
* Look for ways to “bounce back” from any setbacks and grow from any experiences you may be having.
* Have a “next time” mentality - if something doesn’t go your way this time, don’t give up, instead think - “I’ll get it next time around.”

**What Went Well:**

* In order to do quality visualizations, we had to sacrifice a continuous variable and convert it to categorical. Statistically, we wouldn’t want to do that, but it made for a better representation visually.
* Creating the age category ended up giving us some interesting insights as far as how different age groups practice resilience and manage stress.
* R is a great collaborative function, so any stumbling blocks I ran into were easily fixable due to the large platform and helpful users.

**What Did NOT go Well:**

* The animation portion was challenging to implement in R and on the poster.
* Standardizing variables is often a necessity when Likert scaling varies within scales. We were forced to standardize sleep, and it then became apparent that we needed to standardize all of them for specific visualizations.
* R libraries are an awesome collaborative function. However, the downside is they need to be updated and improved along with R. Older packages are better than some newer ones and navigating the old versions and the new versions was complex. Especially for the animation function.

**What Would you do Differently Next Time:**

* I would pick uniform scaling in the initial steps of the survey to avoid z-scoring.
* I would consider a wider range of resilience practices. The study we are doing has 9 resilience practices, and I wish we could have evaluated all of them.

**References**

Bonanno, G. A., & Diminich, E. D. (2013). Annual Research Review: Positive adjustment to adversity–trajectories of minimal–impact resilience and emergent resilience. *Journal of Child Psychology and Psychiatry*, *54*(4), 378-401.

Carver, C. S. (1997). You want to measure coping but your protocol’s too long: Consider the Brief COPE. *International Journal of Behavioral Medicine*, 4, 92-100.

Carver, C. S. (2019). Coping. In C. D. Llewellyn, S. Ayers, C. McManus, S. Newman, K. J. Petrie, T. A. Revenson, & J. Weinman (Eds.), *The Cambridge handbook of psychology, health and medicine., 3rd ed.* (pp. 114–118). Cambridge University Press.

Carver, C. S., Scheier, M. F., & Weintraub, J. K. (1989). Assessing coping strategies: a theoretically based approach. *Journal of Personality and Social Psychology*, *56*(2), 267-283.

Cohen, S., Kamarck, T., and Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24,*386-396.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random House.

Kearns, H., & Gardiner, M. (2007). Is it time well spent? The relationship between time management behaviours, perceived effectiveness and work‐related morale and distress in a university context. *High Education Research & Development*, *26*(2), 235-247.

Marinus, J., Visser, M., van Hilten, J. J., Lammers, G. J., & Stiggelbout, A. M. (2003). Assessment of sleep and sleepiness in Parkinson disease. Sleep, 26(8), 1049-1054.

Parent, M. C. (2013). Handling item-level missing data: {Simpler} is just as good. *The Counseling Psychologist*, *41*(4), 568–600. https://doi.org/10.1177/0011000012445176

Smith, B. W., Dalen, J., Wiggins, K., Tooley, E., Christopher, P., & Bernard, J. (2008). The brief resilience scale: assessing the ability to bounce back. *International Journal of Behavioral Medicine*, *15*(3), 194-200.

Tabibnia, G., & Radecki, D. (2018). Resilience training that can change the brain. *Consulting Psychology Journal: Practice and Research, 70*(1), 59-88.