Final Project NHIS

Matthew Roland

```r
# load data

link <- "https://media.githubusercontent.com/media/Mattr5541/DATA_606_Final_Project/main/adult22.csv"


nhis <- read.csv(url(link), stringsAsFactors = F)

colnames(nhis)
```

```
##   [1] "URBRRL"       "RATCAT_A"     "INCTCFLG_A"   "IMPINCFLG_A"  "SHOTTYPE1_A"
##   [6] "CEVOTELC_A"   "CEMMETNG_A"   "CEVOLUN2_A"   "CEVOLUN1_A"   "HITTEST_A"
##  [11] "HITCOMM_A"    "HITLOOK_A"    "ACCSSHOM_A"   "ACCSSINT_A"   "PPSU"
##  [16] "PSTRAT"       "WLKLEISTC_A"  "WLKTRANTC_A"  "HISPALLP_A"   "RACEALLP_A"
##  [21] "DISAB3_A"     "SCHDYMSSTC_A" "AFNOW"        "PHQCAT_A"     "YRSINUS_A"
##  [26] "CITZNSTP_A"   "PRTNREDUCP_A" "SPOUSEDUCP_A" "LEGMSTAT_A"   "MARSTAT_A"
##  [31] "SASPPRACE_A"  "SASPPHISP_A"  "PRTNRAGETC_A" "SPOUSAGETC_A" "PRTNRWKFT_A"
##  [36] "PRTNRWRK_A"   "SPOUSWKFT_A"  "SPOUSWRK_A"   "SPOUSESEX_A"  "PRTNRSEX_A"
##  [41] "SHOTTYPER_A"  "SHINGYEARP_A" "HHRESPSA_FLG" "GADCAT_A"     "PCNTADTWFP_A"
##  [46] "PCNTADTWKP_A" "FDSCAT4_A"    "FDSCAT3_A"    "EPINUMSEZP_A" "EMPDYSMSS3_A"
##  [51] "EMPLSTWOR1_A" "EMPWRKFT1_A"  "EMPWRKLSW1_A" "EMPWKHRS3_A"  "DIBAGETC_A"
##  [56] "DIFYRSTC1_A"  "SMKQTNP_A"    "SMKQTY_A"     "SMKECIGST_A"  "SMKCIGST_A"
##  [61] "BMICAT_A"     "WEIGHTLBTC_A" "HEIGHTTC_A"   "DRKHVY12M_A"  "DRKSTAT_A"
##  [66] "DRK12MYR_A"   "DRK12MWK_A"   "URGNT12MTC_A" "EMERG12MTC_A" "PA18_05R_A"
##  [71] "PA18_02R_A"   "MODFREQW_A"   "MODNR_A"      "MODTPR_A"     "MODLNR_A"
##  [76] "MODLTPR_A"    "MODMIN_A"     "VIGNR_A"      "VIGTPR_A"     "VIGLNR_A"
##  [81] "VIGLTPR_A"    "VIGFREQW_A"   "VIGMIN_A"     "STRTPR_A"     "STRNR_A"
##  [86] "STRFREQW_A"   "PCNT18UPTC"   "PCNTLT18TC"   "COVER65_A"    "COVER_A"
##  [91] "EXCHANGE_A"   "NOTCOV_A"     "MILSPC1R_A"   "OGFLG_A"      "OPFLG_A"
##  [96] "CHFLG_A"      "MAFLG_A"      "PLNWRKR2_A"   "PLNWRKR1_A"   "RSNHIMISS_A"
## [101] "RSNHIJOB_A"   "MCADVR_A"     "PRFLG_A"      "PLEXCHPR1_A"  "PRPREM1_A"
## [106] "PXCHNG1_A"    "HICOSTR2_A"   "HICOSTR1_A"   "OTHGOV_A"     "OTHPUB_A"
## [111] "IHS_A"        "MILITARY_A"   "CHIP_A"       "MEDICAID_A"   "MEDICARE_A"
## [116] "PRIVATE_A"    "PRPLCOV1_C_A" "PRPLCOV2_C_A" "PLEXCHOG_A"   "PLEXCHOP_A"
## [121] "EXCHPR2_A"    "EXCHPR1_A"    "EDUCP_A"      "MAXEDUCP_A"   "PARSTAT_A"
## [126] "SAPARENTSC_A" "MLTFAMFLG_A"  "OVER65FLG_A"  "PCNTADLT_A"   "PCNTKIDS_A"
## [131] "NUMCAN_A"     "COLRCAGETC_A" "HDNCKAGETC_A" "OTHERAGETC_A" "UTERUAGETC_A"
## [136] "THYROAGETC_A" "THROAAGETC_A" "STOMAAGETC_A" "SKNDKAGETC_A" "SKNNMAGETC_A"
## [141] "SKNMAGETC_A"  "RECTUAGETC_A" "PROSTAGETC_A" "PANCRAGETC_A" "OVARYAGETC_A"
## [146] "MOUTHAGETC_A" "MELANAGETC_A" "LYMPHAGETC_A" "LUNGAGETC_A"  "LIVERAGETC_A"
## [151] "LEUKEAGETC_A" "LARYNAGETC_A" "GALLBAGETC_A" "ESOPHAGETC_A" "COLONAGETC_A"
## [156] "CERVIAGETC_A" "BREASAGETC_A" "BRAINAGETC_A" "BONEAGETC_A"  "BLOODAGETC_A"
## [161] "BLADDAGETC_A" "OTHERCANP_A"  "COLRCCAN_A"   "HDNCKCAN_A"   "UTERUCAN_A"
## [166] "THYROCAN_A"   "THROACAN_A"   "STOMACAN_A"   "SKNDKCAN_A"   "SKNNMCAN_A"
## [171] "SKNMCAN_A"    "RECTUCAN_A"   "PROSTCAN_A"   "PANCRCAN_A"   "OVARYCAN_A"
## [176] "MOUTHCAN_A"   "MELANCAN_A"   "LYMPHCAN_A"   "LUNGCAN_A"    "LIVERCAN_A"
## [181] "LEUKECAN_A"   "LARYNCAN_A"   "GALLBCAN_A"   "ESOPHCAN_A"   "COLONCAN_A"
## [186] "CERVICAN_A"   "BREASCAN_A"   "BRAINCAN_A"   "BONECAN_A"    "BLOODCAN_A"
## [191] "BLADDCAN_A"   "HISDETP_A"    "HISP_A"       "REGION"       "INTV_QRT"
## [196] "SRVY_YR"      "SEX_A"        "AGEP_A"       "AGE65"        "ASTATNEW"
## [201] "TRANSPOR_A"   "HOUGVASST_A"  "HOUTENURE_A"  "HOUYRSLIV_A"  "FDSNEDAYS_A"
## [206] "FDSNOTEAT_A"  "FDSWEIGHT_A"  "FDSHUNGRY_A"  "FDSLESS_A"    "FDSSKIPDYS_A"
## [211] "FDSSKIP_A"    "FDSBALANCE_A" "FDSLAST_A"    "FDSRUNOUT_A"  "FLUNCH12M_A"
## [216] "FWIC12M_A"    "FSNAP30D_A"   "FSNAP12M_A"   "INCOTHR_A"    "INCRETIRE_A"
## [221] "INCWELF_A"    "SSISSDIDSB_A" "SSISSDIBTH_A" "INCSSISSDI_A" "INCSSRR_A"
## [226] "INCINTER_A"   "INCWRKO_A"    "EMPHEALINS_A" "EMPSICKLV_A"  "EMPWHENWRK_A"
## [231] "EMPWHYNOT_A"  "EMPNOWRK_A"   "EMPLASTWK_A"  "SCHCURENR_A"  "NATUSBORN_A"
## [236] "VACAREEV_A"   "VAHOSP_A"     "VADISB_A"     "COMBAT_A"     "AFVETTRN_A"
## [241] "AFVET_A"      "EVRMARRIED_A" "SPOUSEP_A"    "SPOUSLIV_A"   "MARITAL_A"
## [246] "ORIENT_A"     "VSLBTWR_A"    "VSLPALF_A"    "VSLPA_A"      "VSLBRAIN_A"
## [251] "VSLLGDIF_A"   "VSLLGDYS_A"   "VSLLGYR_A"    "VSLSPST_A"    "VSLSPDIF_A"
## [256] "VSLSPDYS_A"   "VSLSPYR_A"    "VSLSWDIF_A"   "VSLSWDYS_A"   "VSLSWYR_A"
## [261] "VSLVDIF_A"    "VSLVDYS_A"    "VSLVYR_A"     "YOGAHLTH_A"   "YOGAPAIN_A"
## [266] "YOGAMED_A"    "YOGABRTH_A"   "YOGA_A"       "GIPRHLTH_A"   "GIPRPAIN_A"
## [271] "GIPR_A"       "MEDIHLTH_A"   "MEDIPAIN_A"   "MEDITATE_A"   "MUSICTHPY_A"
## [276] "ARTTHPY_A"    "NATURHLTH_A"  "NATURPAIN_A"  "NATUR_A"      "MASSHLTH_A"
## [281] "MASSPAIN_A"   "MASS_A"       "ACUHLTH_A"    "ACUPAIN_A"    "ACU_A"
## [286] "CHIROHLTH_A"  "CHIROPAIN_A"  "CHIRO_A"      "TOMSAUTP_A"   "TOMSAUNO_A"
## [291] "PIZZATP_A"    "PIZZANO_A"    "SALSATP_A"    "SALSANO_A"    "OVEGTP_A"
## [296] "OVEGNO_A"     "BEANSTP_A"    "BEANSNO_A"    "POTATOTP_A"   "POTATONO_A"
## [301] "FRIESTP_A"    "FRIESNO_A"    "SALADTP_A"    "SALADNO_A"    "FRUITTP_A"
## [306] "FRUITNO_A"    "FRTDRTP_A"    "FRTDRNO_A"    "SPORDRTP_A"   "SPORDRNO_A"
## [311] "COFFEENOTP_A" "COFFEENO_A"   "FRJUICTP_A"   "FRJUICNO_A"   "SODATP_A"
## [316] "SODANO_A"     "SLPMED_A"     "SLPSTY_A"     "SLPFLL_A"     "SLPREST_A"
## [321] "SLPHOURS_A"   "ADVACTIVE_A"  "WLKLEISTPD_A" "WLKLEISDAY_A" "WLKLEIS_A"
## [326] "WLKTRANTPD_A" "WLKTRANDAY_A" "WLKTRAN_A"    "DRKADVISE1_A" "DRKBNG30D_A"
## [331] "DRKANY30D_A"  "DRKBNG12M_A"  "DRK12ANYR_A"  "DRKAVG12M_A"  "DRK12MTP_A"
## [336] "DRK12MN_A"    "DRKLIFE_A"    "HPTOB3_A"     "SMOKELSCUR_A" "SMOKELSEV_A"
## [341] "PIPECUR_A"    "PIPEEV_A"     "CIGAR30D_A"   "CIGARCUR_A"   "CIGAREV_A"
## [346] "QWANT_A"      "CQUITE_A"     "CQUITB3_A"    "CQUITB2_A"    "CQUITB1_A"
## [351] "CQUITA5_A"    "CQUITA4_A"    "CQUITA3_A"    "CQUITA2_A"    "CQUITA1_A"
## [356] "FQUITE_A"     "FQUITB3_A"    "FQUITB2_A"    "FQUITB1_A"    "FQUITA5_A"
## [361] "FQUITA4_A"    "FQUITA3_A"    "FQUITA2_A"    "FQUITA1_A"    "ECIGNOW_A"
## [366] "ECIGEV_A"     "SMKTLK1_A"    "MENTHOLF_A"   "SMKQTTP_A"    "SMKQT12M_A"
## [371] "MENTHOLC_A"   "CIG30D_A"     "SMK30D_A"     "CIGNOW_A"     "SMKNOW_A"
## [376] "SMKAGE_A"     "SMKEV_A"      "FGELEVTRD_A"  "FGELNGTRD_A"  "FGEFRQTRD_A"
## [381] "GAD77_A"      "GAD76_A"      "GAD75_A"      "GAD74_A"      "GAD73_A"
## [386] "GAD72_A"      "GAD71_A"      "PHQ88_A"      "PHQ87_A"      "PHQ86_A"
## [391] "PHQ85_A"      "PHQ84_A"      "PHQ83_A"      "PHQ82_A"      "PHQ81_A"
## [396] "MHTHND_A"     "MHTHDLY_A"    "MHTPYNOW_A"   "MHTHRPY_A"    "MHRX_A"
## [401] "DEPLEVEL_A"   "DEPMED_A"     "DEPFREQ_A"    "ANXLEVEL_A"   "ANXMED_A"
## [406] "ANXFREQ_A"    "HOMEHC12M_A"  "THERA12M_A"   "EYEEX12M_A"   "WRKHLTHFC_A"
## [411] "WORKHEALTH_A" "SHTHPVAGE_A"  "SHTHPV_A"     "SHTTDAP_A"    "SHTTETANUS_A"
## [416] "TDAPPREG_A"   "SHINGRIXFS_A" "SHINGRIXN2_A" "SHINGRIX2_A"  "SHINGWHEN1_A"
## [421] "SHTSHINGL1_A" "SHTPNEUNB_A"  "SHTPNUEV_A"   "SHOTTYPE_A"   "CVDVAC2Y_A"
## [426] "CVDVAC2M_A"   "CVDVAC1Y_A"   "CVDVAC1M_A"   "SHTCVD19NM_A" "SHTCVD191_A"
## [431] "FLUPREG2_A"   "FLUPREG_A"    "SHTFLUY_A"    "SHTFLUM_A"    "SHTFLU12M_A"
## [436] "LIVEBIRTH_A"  "PREGFLUYR_A"  "RXDG12M_A"    "RXDL12M_A"    "RXLS12M_A"
## [441] "RXSK12M_A"    "RX12M_A"      "VIRAPP12M_A"  "ABTIME_A"     "ABTOOLONG_A"
## [446] "ABOPEN_A"     "ABAVAIL_A"    "ABINSUR_A"    "MEDNG12M_A"   "MEDDL12M_A"
## [451] "HOSPONGT_A"   "USPLKIND_A"   "USUALPL_A"    "WELLVIS_A"    "WELLNESS_A"
## [456] "LASTDR_A"     "DENNG12M_A"   "DENDL12M_A"   "DENPREV_A"    "SYMPNOW_A"
## [461] "LONGCVD_A"    "CVDSEV_A"     "POSTEST_A"    "CVDDIAG_A"    "PAYWORRY_A"
## [466] "PAYNOBLLNW_A" "PAYBLL12M_A"  "HINOTMYR_A"   "HINOTYR_A"    "RSNHIOTH_A"
## [471] "RSNHIWAIT_A"  "RSNHIMEET_A"  "RSNHICONF_A"  "RSNHIELIG_A"  "RSNHIWANT_A"
## [476] "RSNHICOST_A"  "HISTOPELIG_A" "HISTOPCOST_A" "HISTOPAGE_A"  "HISTOPMISS_A"
## [481] "HISTOPJOB_A"  "HILASTMY_A"   "HILAST_A"     "MILSPC3_A"    "MILSPC2_A"
## [486] "MILSPC1_A"    "OGHDHP_A"     "OGDEDUC_A"    "OGPREM_A"     "OGXCHNG_A"
## [491] "OPHDHP_A"     "OPDEDUC_A"    "OPPREM_A"     "OPXCHNG_A"    "CHHDHP_A"
## [496] "CHDEDUC_A"    "CHPREM_A"     "CHXCHNG_A"    "PRVSCOV2_A"   "PRVSCOV1_A"
## [501] "PRDNCOV2_A"   "PRDNCOV1_A"   "PRRXCOV2_A"   "PRRXCOV1_A"   "HSAHRA2_A"
## [506] "HSAHRA1_A"    "PRHDHP2_A"    "PRHDHP1_A"    "PRDEDUC2_A"   "PRDEDUC1_A"
## [511] "PLN2PAY6_A"   "PLN2PAY5_A"   "PLN2PAY4_A"   "PLN2PAY3_A"   "PLN2PAY2_A"
## [516] "PLN2PAY1_A"   "PLN1PAY6_A"   "PLN1PAY5_A"   "PLN1PAY4_A"   "PLN1PAY3_A"
## [521] "PLN1PAY2_A"   "PLN1PAY1_A"   "PLNEXCHG2_A"  "PLNEXCHG1_A"  "PRPOLH2_A"
## [526] "PRPOLH1_A"    "PRPLCOV2_A"   "PRPLCOV1_A"   "POLHLD2_A"    "POLHLD1_A"
## [531] "MAHDHP_A"     "MADEDUC_A"    "MAPREM_A"     "MAXCHNG_A"    "MCPARTD_A"
## [536] "MCVSCOV_A"    "MCDNCOV_A"    "MCHMO_A"      "MCCHOICE_A"   "MCPART_A"
## [541] "SINCOVRX_A"   "SINCOVVS_A"   "SINCOVDE_A"   "MCAIDPRB_A"   "MCAREPRB_A"
## [546] "HIKIND10_A"   "HIKIND09_A"   "HIKIND08_A"   "HIKIND07_A"   "HIKIND06_A"
## [551] "HIKIND05_A"   "HIKIND04_A"   "HIKIND03_A"   "HIKIND02_A"   "HIKIND01_A"
## [556] "HICOV_A"      "DEVDONSET_A"  "SOCWRKLIM_A"  "SOCSCLPAR_A"  "SOCERRNDS_A"
## [561] "UPPOBJCT_A"   "UPPRAISE_A"   "UPPSLFCR_A"   "COGAMTDFF_A"  "COGFRQDFF_A"
## [566] "COGTYPEDFF_A" "COGMEMDFF_A"  "COMDIFF_A"    "EQSTEPS_A"    "EQWLK13M_A"
## [571] "EQWLK100_A"   "NOEQSTEPS_A"  "NOEQWLK13M_A" "NOEQWLK100_A" "PERASST_A"
## [576] "WCHAIR_A"     "CANEWLKR_A"   "STEPS_A"      "WLK13M_A"     "WLK100_A"
## [581] "EQUIP_A"      "DIFF_A"       "HEARINGDF_A"  "HEARAIDFR_A"  "HEARAID_A"
## [586] "VISIONDF_A"   "WEARGLSS_A"   "PREGNOW_A"    "EPIDR_A"      "EPIMED_A"
## [591] "EPIEV_A"      "HLTHCOND_A"   "MEDRXTRT_A"   "CFSNOW_A"     "CFSEV_A"
## [596] "DEPEV_A"      "ANXEV_A"      "DEMENEV_A"    "ARTHEV_A"     "COPDEV_A"
## [601] "DIBTYPE_A"    "DIBINSSTYR_A" "DIBINSSTOP_A" "DIBINSTIME_A" "DIBINS_A"
## [606] "DIBPILL_A"    "DIBEV_A"      "GESDIB_A"     "PREDIB_A"     "CANEV_A"
## [611] "ASER12M_A"    "ASAT12M_A"    "ASTILL_A"     "ASEV_A"       "STREV_A"
## [616] "MIEV_A"       "ANGEV_A"      "CHDEV_A"      "CHLMED_A"     "CHL12M_A"
## [621] "CHLEV_A"      "HYPMED_A"     "HYP12M_A"     "HYPDIF_A"     "HYPEV_A"
## [626] "LSATIS4_A"    "PHSTAT_A"     "PROXYREL_A"   "PROXY_A"      "AVAIL_A"
## [631] "HHSTAT_A"     "INTV_MON"     "RECTYPE"      "IMPNUM_A"     "WTFA_A"
## [636] "HHX"          "POVRATTC_A"
```

```r
##I will subset my variables of interest, including participant IDs, race, region (metropolitan vs. non

nhis_vars <- nhis %>% subset(select = c(HHX, RACEALLP_A, URBRRL, SEX_A, AGEP_A, MHRX_A, MHTHRPY_A, MHTH
```

Abstract

This analysis aims to examine the relationships among region, educational attainment, and depression levels, as indicated by the PHQ scale. I hypothesized that: 1) living in metropolitan regions would be associated with higher levels of depression; 2) having higher levels of educational attainment would predict lower levels of depression; 3) educational achievement would moderate the relationship between region and depression such that individuals in non-metropolitan regions would have lower levels of depression. A multiple regression analysis was used to assess this relationship, using mean PHQ scores as the outcome variable, educational attainment as a predictor, and region as a nominal predictor. In contrast to expectations, living in non-

metropolitan regions was associated with higher levels of depression than living in metropolitan regions. Also, higher levels of education were associated with higher levels of depression. However, an interaction effect revealed that educational achievement moderated mean depression scores such that individuals living in smaller metropolitan areas or non-metropolitan areas were more likely to have lower levels of depression when educational attainment was higher (Undergraduate to graduate/post-graduate). Although not entirely in line with expectations, these results may have implications regarding the importance of providing mental health care for individuals in non-metropolitan regions. Also, the results may have implications for the role of education in moderating depression in smaller metropolitan regions. Of course, none of the findings provided in this analysis are causal in nature.

Part 1 - Introduction

An important consideration in our modern, densely populated world is the impact of living conditions and residence on our mental health. Such topics have been studied indirectly via animal research and through observational approaches in humans. An example of this would type of research would be John Calhoon's seminal "Behavioral Sink" experiment, in which rats were placed into overcrowded conditions to examine the effects of living condition on behavior (Calhoon, 1962). Researchers have attempted to generalize the results of Calhoon's work to humans, specifically, within the context of how crowded yet socially isolated urban environments may impact mental health. Of course, due to the nature of this type of research question, experimental approaches are severely limited; thus, most of the evidence that we have gathered is through observational and archival approaches, which cannot provide causal evidence of the impacts of such environments on mental health. Typically, researchers have found that urban environments have higher incidences of poor mental health outcomes (cite, if possible)

Another question that individuals may pose would be the link between education and mental health outcomes. Researchers have found that higher educational achievements corresponds with better mental health outcomes in adults (Bauldry, 2015)

My goal is to expand on these findings by examining the relationships among environmental conditions, education, and mental health. Specifically, I want to know whether living in a metropolitan region (large or small) versus a smaller, non-metropolitan region is associated with higher rates of depression. In addition, I am interested in seeing whether lower levels of one's highest education are associated with higher rates of depression. Finally, I want to see if there is an interaction effect between region and education on rates of depression.

The data used in this analysis were sourced from the 2022 National Health Interview Survey (NHIS) data collected by the CDC. This dataset contains survey and questionnaire data detailing demographic and mental/physical health data for individuals in the US in the year 2022. I will be using the adult survey dataset, and I plan to restrict my analysis to adults aged 25 - 36, as I am interested in seeing the environmental, education, and depression trends in the late Gen Z and Millennial generations.

Variable Information

Region Region classifications were defined along CDC guidelines. **large central metro regions** were defined as counties in metropolitan areas of one million or more individuals and either: contain the largest principle city in the metro area; are contained in the largest principle city; or contain at least 250,000 residents of any principle city of the metro area. **Large fringe metro areas** were defined as counties in metropolitan areas of one million or more individuals that do not possess the previously listed criteria that define a large central metro region. **Medium and small metro regions** are defined as counties that in metropolitan areas that contain 250,000 to 999,000 individuals, or metropolitan areas that contain less than 250,000 individuals. Finally, **nonmetropolitan regions** are defined as counties in micropolitan areas and non-core counties

The information provided above were derived from the NHIS codebook, and more detail can be found at the CDC's website: https://www.cdc.gov/nchs/data_access/urban_rural.htm.

Education Education level was assessed via self-reports. Participants were asked about their highest levels of education, which ranged from “never attended/kindergarten only” to “Professional School or Doctoral degree”, along with “Don’t Know” and “Not Ascertained” options. For the sake of simplicity, I will only be analyzing a subset of these selections, including: “Grade 1-11,” “High School Graduate,” “Bachelor’s degree”, and I will combine both “Master’s degree” and “Professional School or Doctoral degree” into a single metric

Depression Depression was assessed via the 8-item Patient Health Questionnaire (PHQ-8; Kroenke et al., 2009), which measures symptoms associated with depression. This inventory contains items that correspond with DSM-IV depression criteria; however, the 8-item scale excludes a question on the 9-question scale pertaining to thoughts pertaining to self-harm and death. This dataset includes responses to all eight items, as well as a metric that aggregates each individuals’ scores across each item to assess their levels of depression (this is represented by the PHQCAT_A variable). For my analysis, I will calculate the mean of each participant’s overall PHQ scores and use that as my dependent variable.

Goals & Hypotheses This analysis will explore the relationship among region, education level, depressive symptoms, and the interactions among these variables. To accomplish this, I will perform a Factorial ANOVA using both region and education level as predictors, and depressive symptoms as an outcome. I believe this analysis will be the most appropriate course of action, as my predictors are all categorical in nature, whereas my outcome is an aggregate of scores measured on a Likert scale. While there is some debate as to how Likert measurements should be calculated, it is common-practice to consider them as continuous or quantitative measurements, which makes them conducive to t-tests and ANOVAs as outcome variables.

My hypotheses are as follows:

1. I predict that one’s region of residence will significantly predict levels of depression such that residing in more metropolitan areas will predict higher levels of depression

2. I predict that one’s level of education will significantly predict levels of depression such that higher level of education (beyond high school) will predict higher levels of depression

3. I predict that there will be a significant interaction between region and level of education on levels of depression such that higher levels of education will reduce the effects of region on levels of depression

Part 2 - Data

First, I will prepare my data:

```
##This code will filter the dataset such that only the observations of those between the age range of 2
nhis_vars <- nhis_vars %>% filter(AGEP_A >= 26 & AGEP_A <= 36)

nhis_vars %>% count()
```

Filtering Ages

```
##      n
## 1 4558
```

```
nhis_vars %>% count(AGEP_A) %>% mutate(prop = n / sum(n))
```

```
##    AGEP_A   n       prop
## 1      26 351 0.07700746
## 2      27 369 0.08095656
## 3      28 359 0.07876262
## 4      29 401 0.08797718
## 5      30 439 0.09631417
## 6      31 451 0.09894691
## 7      32 423 0.09280386
## 8      33 420 0.09214568
## 9      34 461 0.10114085
## 10     35 423 0.09280386
## 11     36 461 0.10114085
```

Now, we are left with 4558 remaining observations in this dataset And, as we can see, the age distribution is rather uniform, overall, with around 300 to 500 participants per age

Education **Now I will recode the education variables to better fit my analysis goals**

```
unique(nhis_vars$EDUCP_A)
```

```
##  [1]  8  4 10  9  2  5  1  7 97  6  3 99
```

```
#Note, I will only be using the following codes: 01 = Grade 1 - 11, 04 = High School Graduate, 08 = Bac

#I will also recode some of the other observation codes so they follow this sequence: 1 = Grade 1 - 11,

nhis_vars <- nhis_vars %>% mutate(EDUCP_A = ifelse(EDUCP_A == 4, 2, EDUCP_A))
nhis_vars <- nhis_vars %>% mutate(EDUCP_A = ifelse(EDUCP_A == 8, 3, EDUCP_A))
nhis_vars <- nhis_vars %>% mutate(EDUCP_A = ifelse(EDUCP_A == 9 | EDUCP_A == 10, 4, EDUCP_A))
```

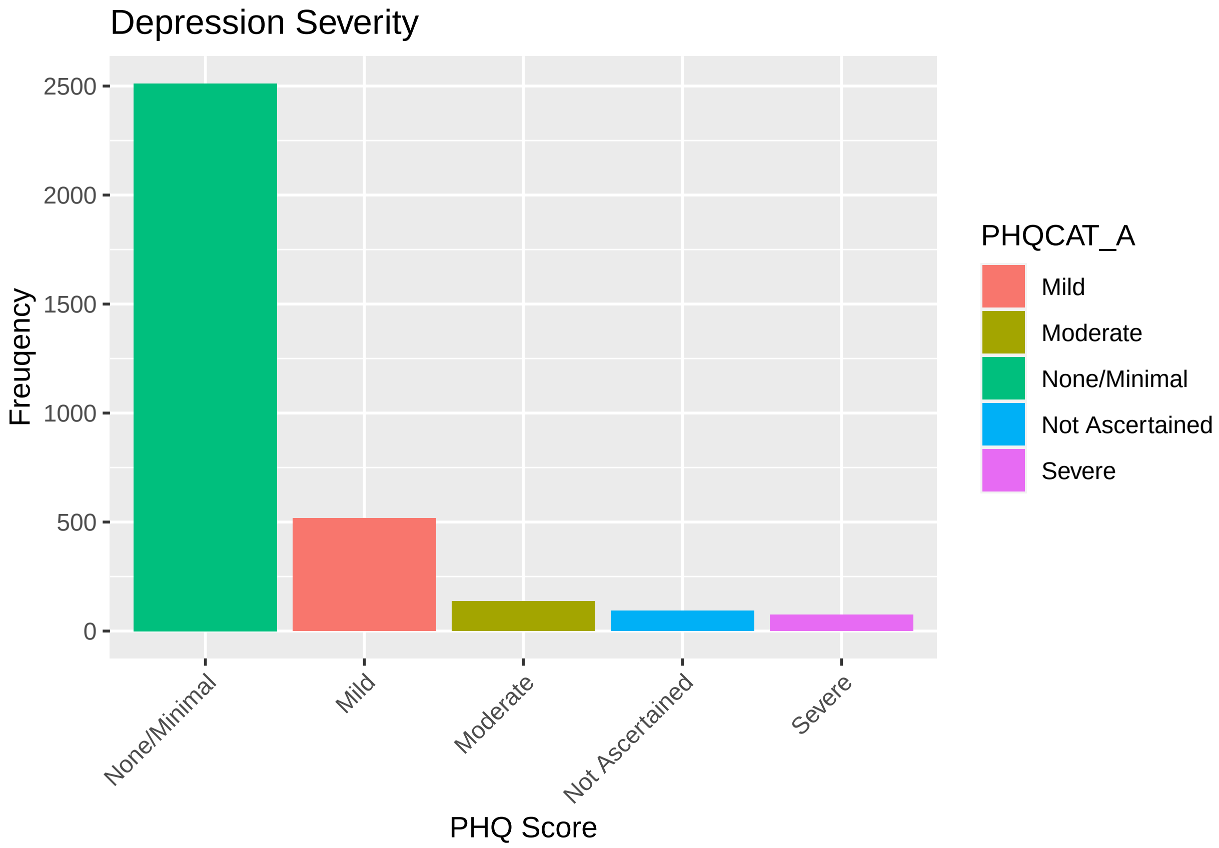
```
nhis_vars <- nhis_vars %>% filter(EDUCP_A < 5)
```

Depression Levels **And finally, let's look at the the frequency of depression levels in the current sample**

```
#The PHQ is a measure of depression. The PHQCAT_A aggregates depression scores on each subscale of the
#Note: 1 = None/Minimal, 2 = Mild, 3 = Moderate, 4 = Severe, 8 = Not Ascertained
nhis_dep <- nhis_vars %>% count(PHQCAT_A) %>% mutate(prop = n / sum(n))

nhis_dep <- nhis_dep %>%
  mutate(PHQCAT_A =
           dplyr::recode(PHQCAT_A,
                  "1" = "None/Minimal",
                  "2" = "Mild",
                  "3" = "Moderate",
                  "4" = "Severe",
                  "8" = "Not Ascertained"))

kable(nhis_dep)
```

PHQCAT_A	n	prop
None/Minimal	2512	0.7536754
Mild	518	0.1554155
Moderate	136	0.0408041
Severe	74	0.0222022
Not Ascertained	93	0.0279028

```
ggplot(nhis_dep, aes(reorder(PHQCAT_A, -n), n, fill = PHQCAT_A)) + geom_bar(stat = "identity")  +  labs
```



```
#I would also like to make a mean variable for each PHQ scale, but first, I need to remove the "Not Asc

nhis_vars <- nhis_vars %>% filter(across(c(PHQCAT_A,PHQ81_A, PHQ82_A, PHQ83_A, PHQ84_A, PHQ85_A, PHQ86_
```

```
## Warning: Using ‘across()‘ in ‘filter()‘ was deprecated in dplyr 1.0.8.
## i Please use ‘if_any()‘ or ‘if_all()‘ instead.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
#And now, I can calculate the overall PHQ score mean for each remaining individual
nhis_mean <- nhis_vars %>% mutate(PHQ_mean = rowMeans(dplyr::select(., PHQ81_A, PHQ82_A, PHQ83_A, PHQ84_

kable(mean(nhis_mean$PHQ_mean))
```

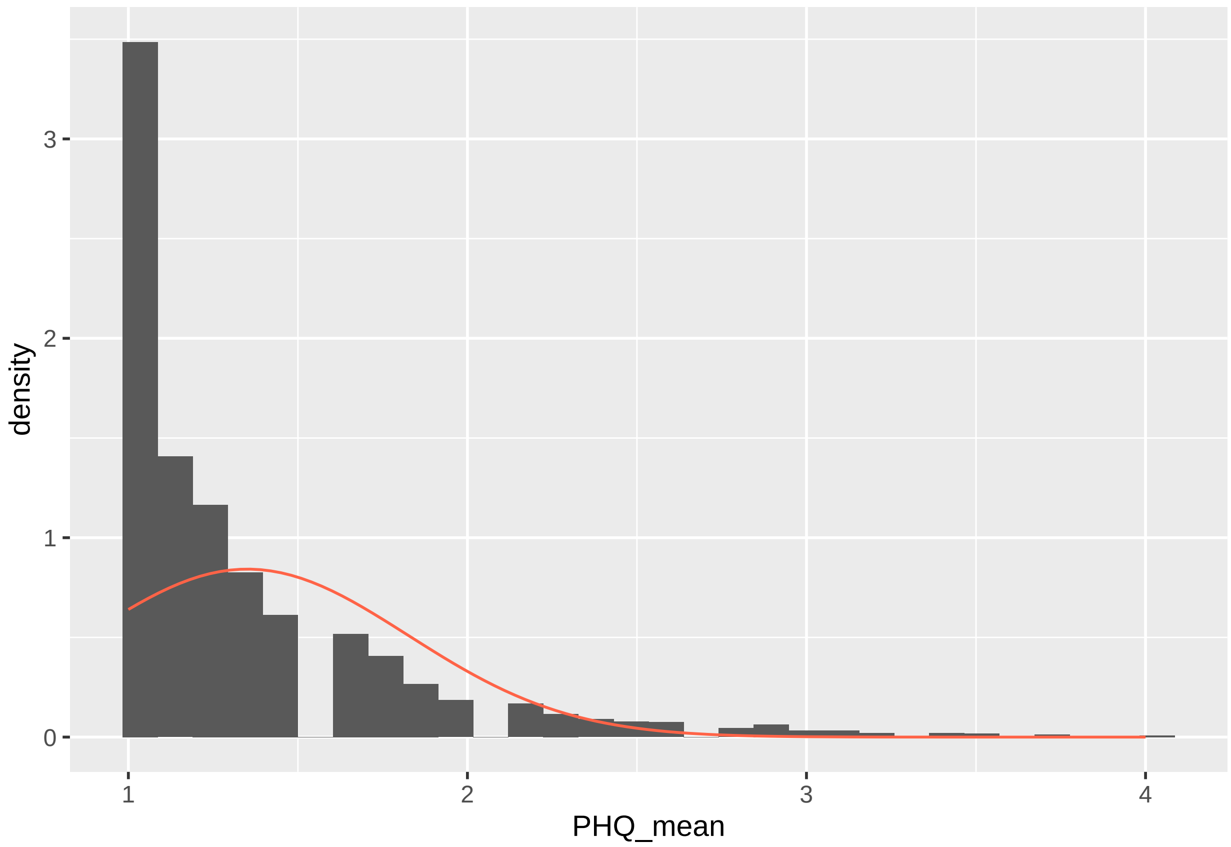
x
1.351084

As one would expect from a distribution representative of the population, most individuals sampled were identified as having mild or no depressive symptoms, with only a few marginal cases having severe depressive symptoms. As a result of this discrepancy, the analyses to come may be significantly impacted. Furthermore, the overall mean of the calculated mean of all PHQ variables is 1.35, when each individual mean can range from 1 - 4 in this dataset. [Therefore, I will run two versions of the analysis, one that included the none/mild cases, and one that includes individuals who are marked as having mild to severe symptoms]

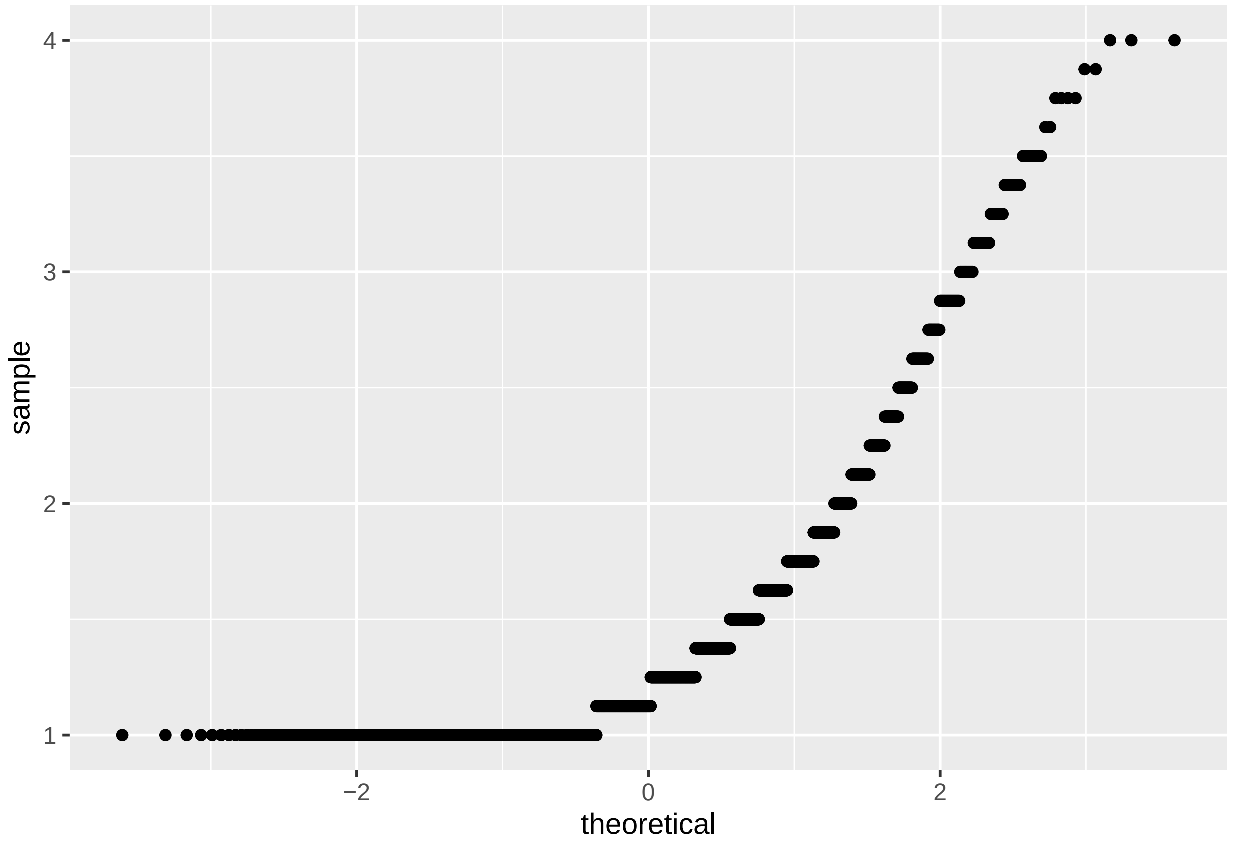
Now let's check the normality of the data

```
ggplot(data = nhis_mean, aes(x = PHQ_mean)) +
        geom_blank() +
        geom_histogram(aes(y = ..density..)) +
        stat_function(fun = dnorm, args = c(mean = mean(nhis_mean$PHQ_mean), sd = sd(nhis_mean$PHQ_mean
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(data = nhis_mean, aes(sample = PHQ_mean)) +
  geom_point(stat = "qq")
```



These graphs indicate that there is a large amount of skew present in the dataset as a result of only a minor portion of the sample having moderate to severe depression symptoms. However, I can attempt to rectifiy this via power transformations

Summary tables and figures **I will now present tables and figures to represent our remaining sample following the previous modifications**

```
nhis_mean %>% count()
```

Age and Gender

```
##      n
## 1 3230
```

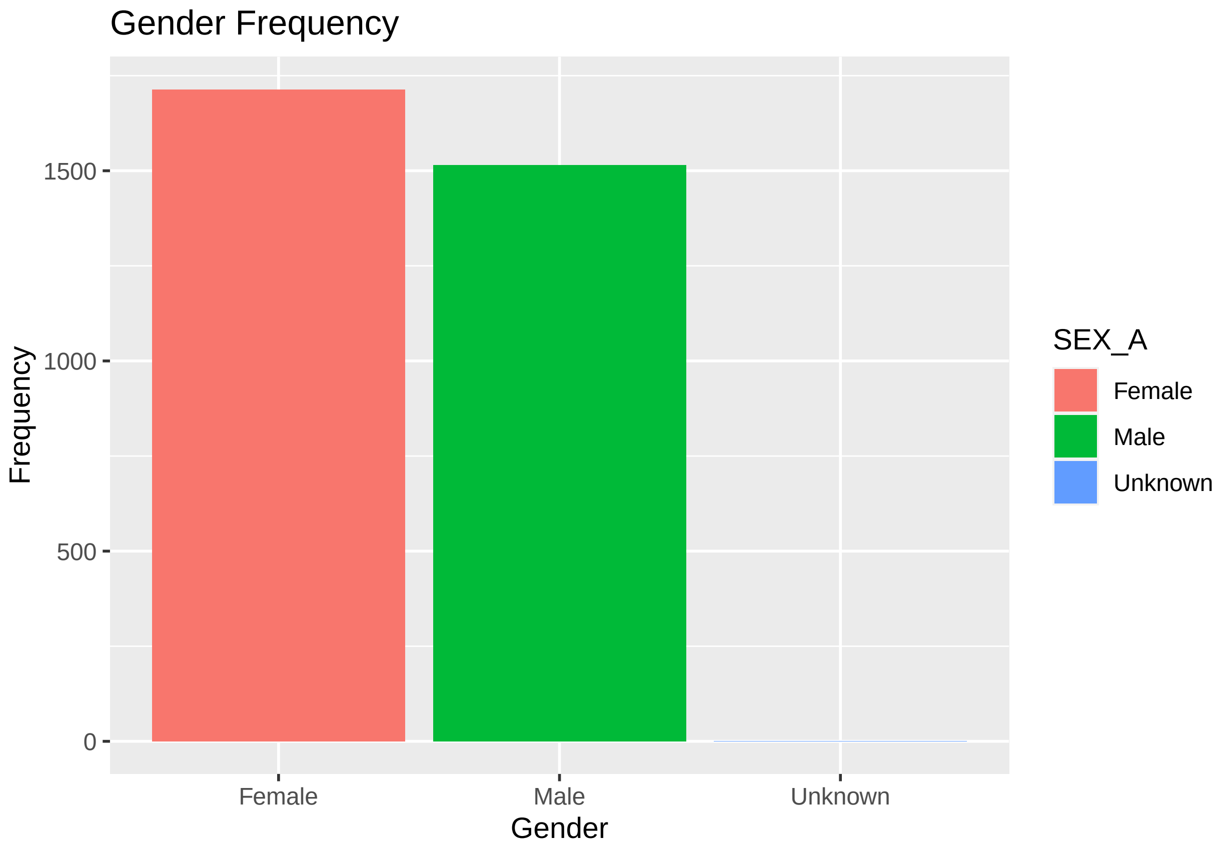
```
nhis_mean %>% count(AGEP_A) %>% mutate(prop = n / sum(n))
```

```
##    AGEP_A   n       prop
## 1      26 249 0.07708978
## 2      27 268 0.08297214
## 3      28 252 0.07801858
## 4      29 290 0.08978328
## 5      30 318 0.09845201
## 6      31 317 0.09814241
## 7      32 296 0.09164087
## 8      33 318 0.09845201
## 9      34 299 0.09256966
## 10     35 302 0.09349845
## 11     36 321 0.09938080
```

```r
nhis_gender <- nhis_mean %>% count(SEX_A) %>% mutate(prop = n / sum(n))

kable(nhis_gender)
```

SEX_A	n	prop
1	1515	0.4690402
2	1714	0.5306502
9	1	0.0003096

```r
nhis_gender <- nhis_gender %>% mutate(SEX_A = dplyr::recode(SEX_A, "1" = "Male", "2" = "Female", "9" = 

ggplot(nhis_gender, aes(SEX_A, n, fill = SEX_A)) + geom_bar(stat = "identity")  +  labs(title = "Gender
```



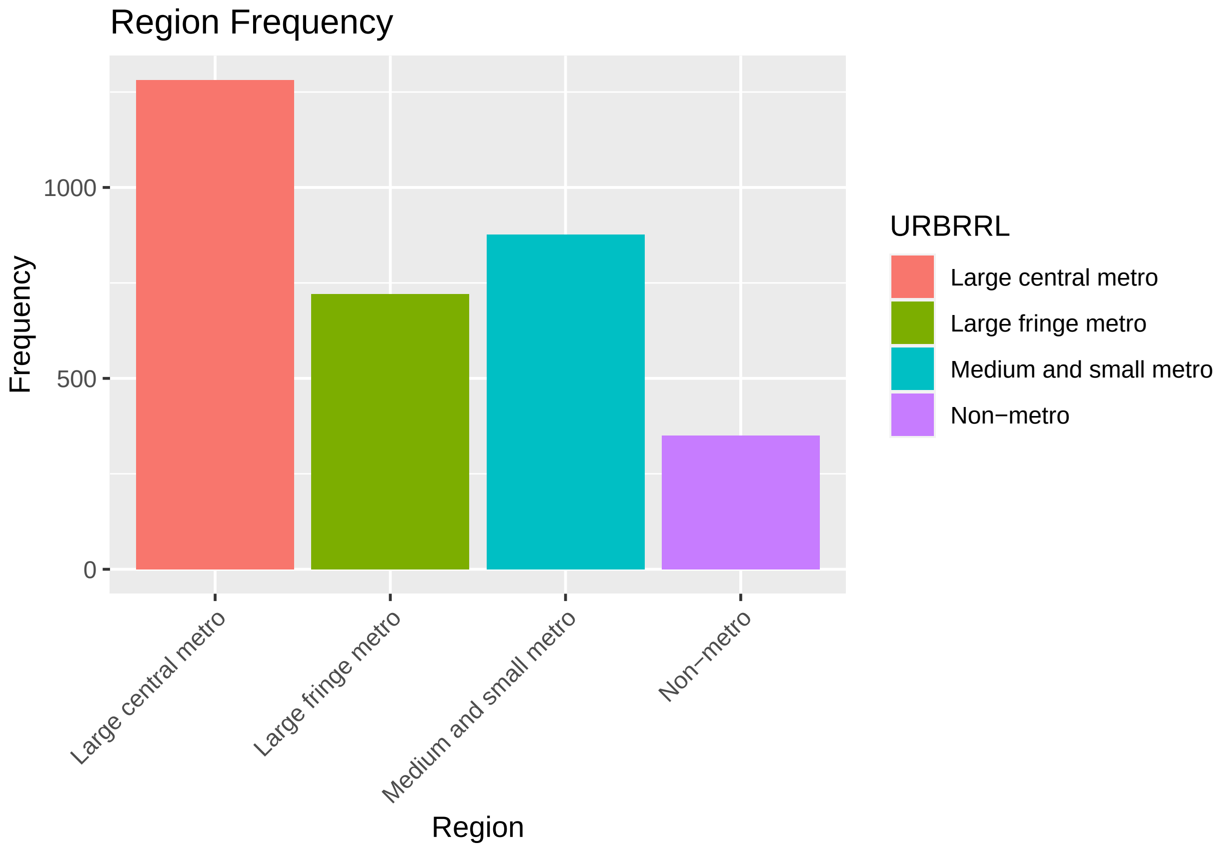
The gender ratio is rather equivalent, with a slight difference such that there are more women than men

```r
# Note: 1 = large central metro, 2 = large fringe metro, 3 = medium and small metro, and 4 = non-metrop
nhis_region <- nhis_mean %>% count(URBRRL) %>% mutate(prop = n / sum(n))

nhis_region <- nhis_region %>%
  mutate(URBRRL =
           dplyr::recode(URBRRL,
                  "1" = "Large central metro",
                  "2" = "Large fringe metro",
                  "3" = "Medium and small metro",
                  "4" = "Non-metro"))

kable(nhis_region)
```

Region

URBRRL	n	prop
Large central metro	1281	0.3965944
Large fringe metro	721	0.2232198
Medium and small metro	877	0.2715170
Non-metro	351	0.1086687

```r
ggplot(nhis_region, aes(URBRRL, n, fill = URBRRL)) + geom_bar(stat = "identity")  +  labs(title = "Regi
```



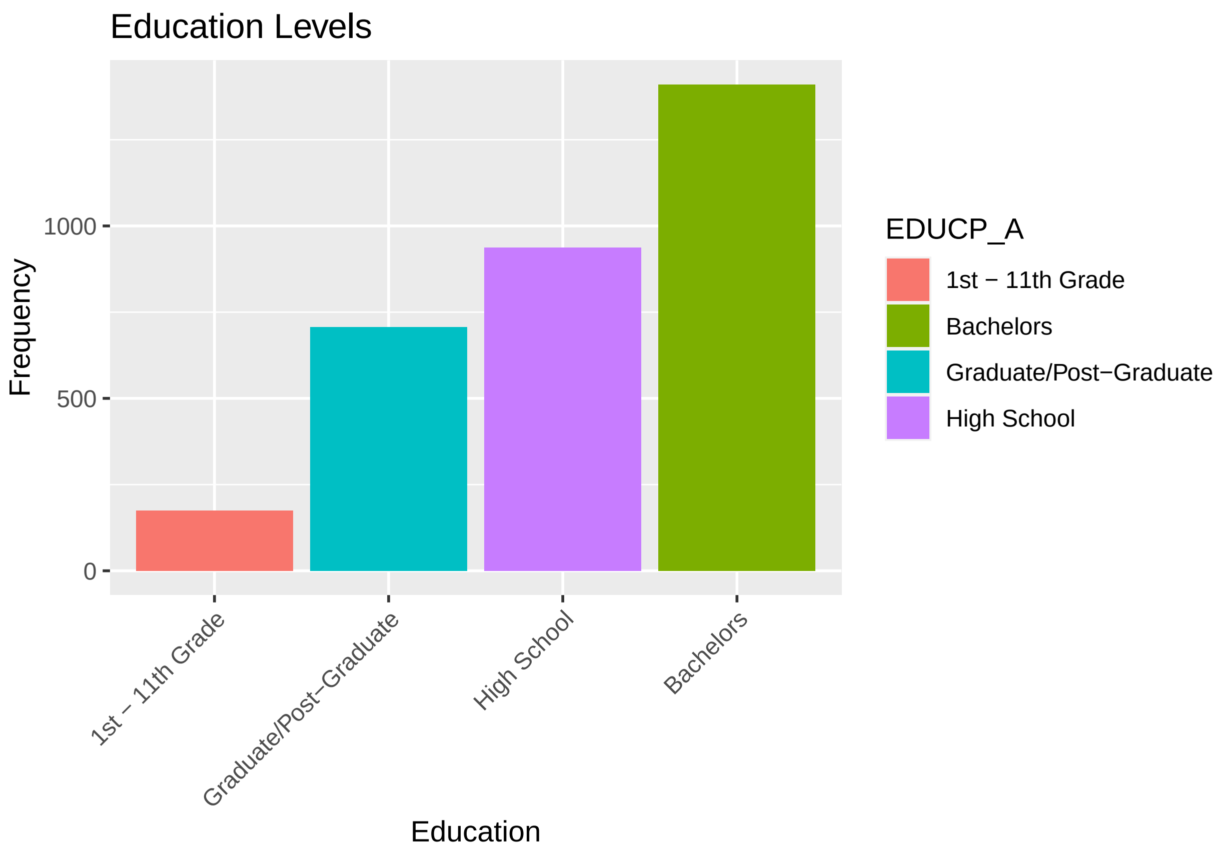
Most of the participants were drawn from large central metro regions, followed by medium and small metro regions, followed by large fringe, and finally, non-metro regions. The disparity among these groups may have potential consequences for the analysis to come

Education

```
nhis_edu <- nhis_vars %>% count(EDUCP_A) %>% mutate(prop = n / sum(n))

nhis_edu <- nhis_edu %>% filter(EDUCP_A %in% c(1, 2, 3, 4, 97, 99))

nhis_edu <- nhis_edu %>% mutate(EDUCP_A = dplyr::recode(EDUCP_A, "1" = "1st - 11th Grade", "2" = "High

kable(nhis_edu)
```

EDUCP_A	n	prop
1st - 11th Grade	175	0.0541796
High School	938	0.2904025
Bachelors	1410	0.4365325
Graduate/Post-Graduate	707	0.2188854

```
ggplot(nhis_edu, aes(reorder(EDUCP_A, n), n, fill = EDUCP_A)) + geom_bar(stat = "identity")  +  labs(ti
```



Most of the participants in this sample have attained at least a high school diploma, with the highest proportion having attained a Bachelor's degree Because of the negligible proportion of individuals who refused to answer or did not know their education level, I will remove those observations from the analysis

Dummy Coding Variables

Because I will be performing a multiple regression analysis, I will need to dummy code my categorical variables

```
nhis_fac <- nhis_mean %>% mutate_at(vars(EDUCP_A, URBRRL), as.factor)

dummy <- model.matrix(~ URBRRL - 1, data = nhis_fac)
nhis_dummy <- nhis_fac %>% cbind(dummy)

dummy <- model.matrix(~ EDUCP_A - 1, data = nhis_fac)
nhis_dummy <- nhis_dummy %>% cbind(dummy)

nhis_dummy <- nhis_dummy %>% dplyr::rename("l_central" ="URBRRL1",
                          "l_fringe" = "URBRRL2",
                          "ms_metro" = "URBRRL3",
                          "non_metro" = "URBRRL4",
                          "drop" = "EDUCP_A1",
                          "high" = "EDUCP_A2",
                          "bachelor" = "EDUCP_A3",
                          "grad" = "EDUCP_A4")
```

Part 3 - Exploratory data analysis

```
summary(p1 <- powerTransform(PHQ_mean ~ (l_fringe + ms_metro + non_metro) * (high + bachelor + grad), da
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -2.2414       -2.24      -2.3901      -2.0927
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df       pval
## LR test, lambda = (0) 1152.006  1 < 2.22e-16
##
## Likelihood ratio test that no transformation is needed
##                            LRT df       pval
## LR test, lambda = (1) 2729.301  1 < 2.22e-16
```

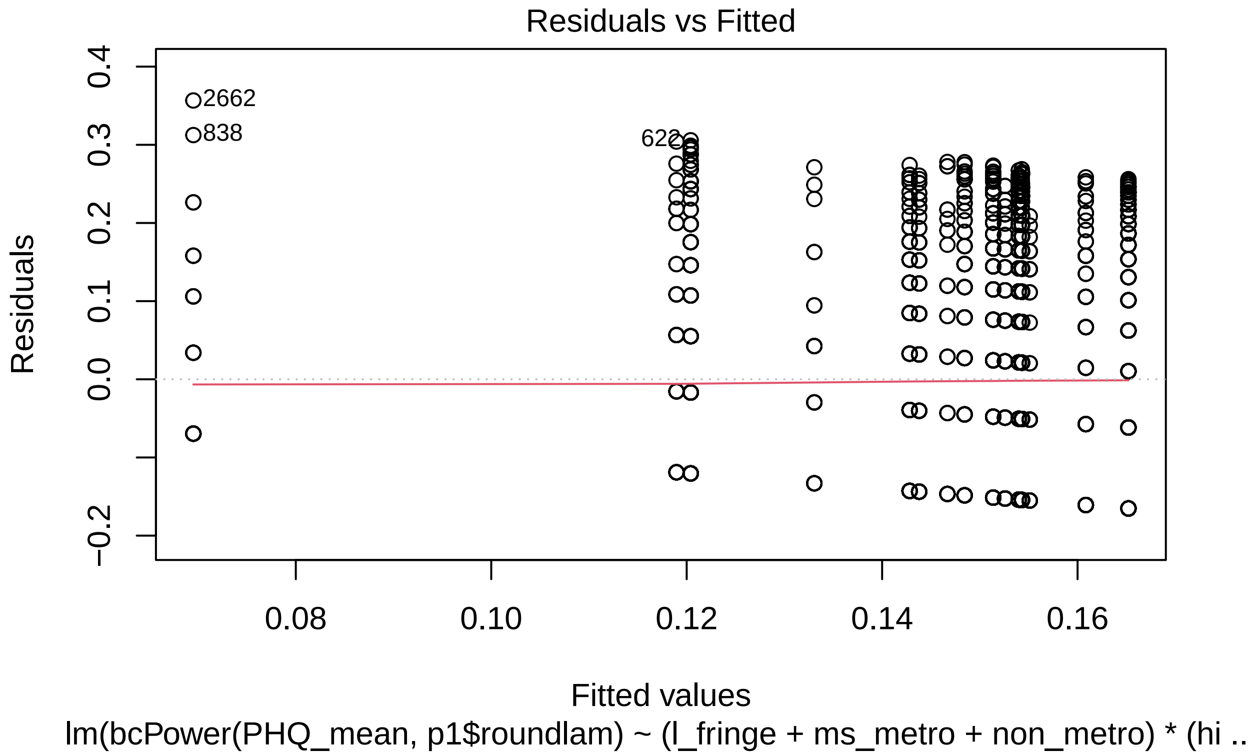
```
coef(p1, round = T)
```

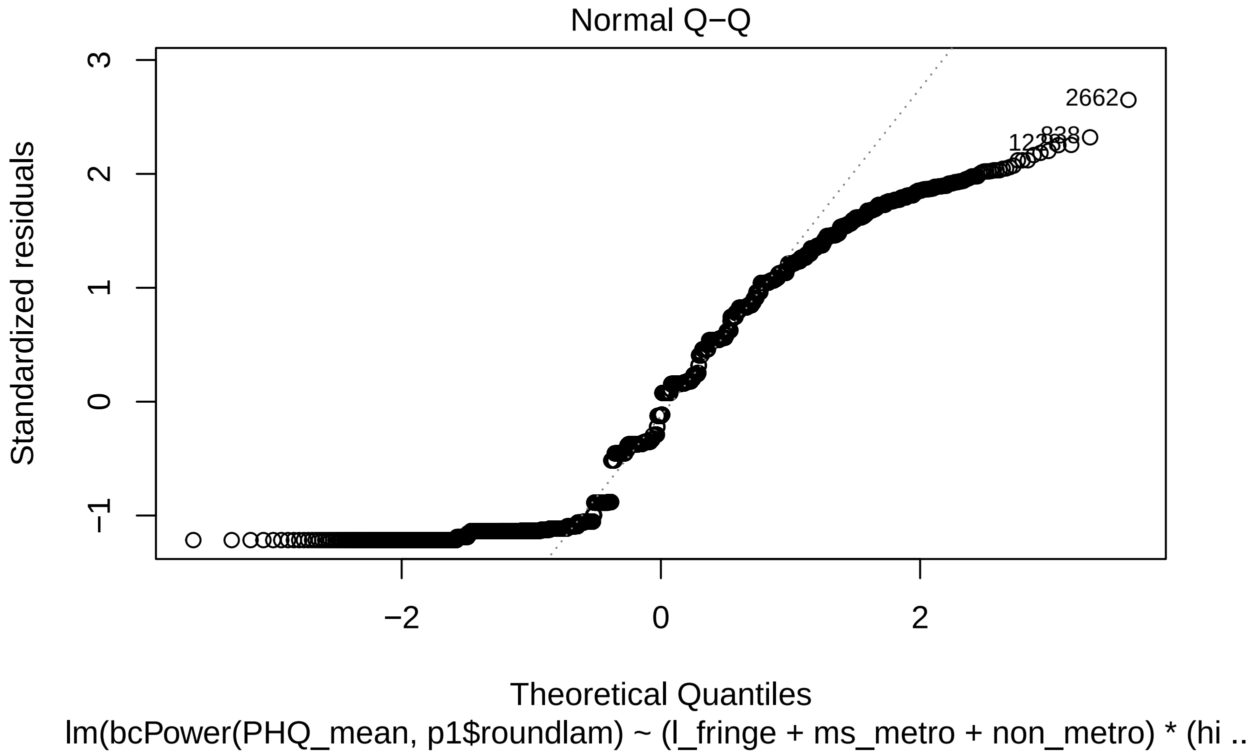
```
##        Y1
## -2.241431
```

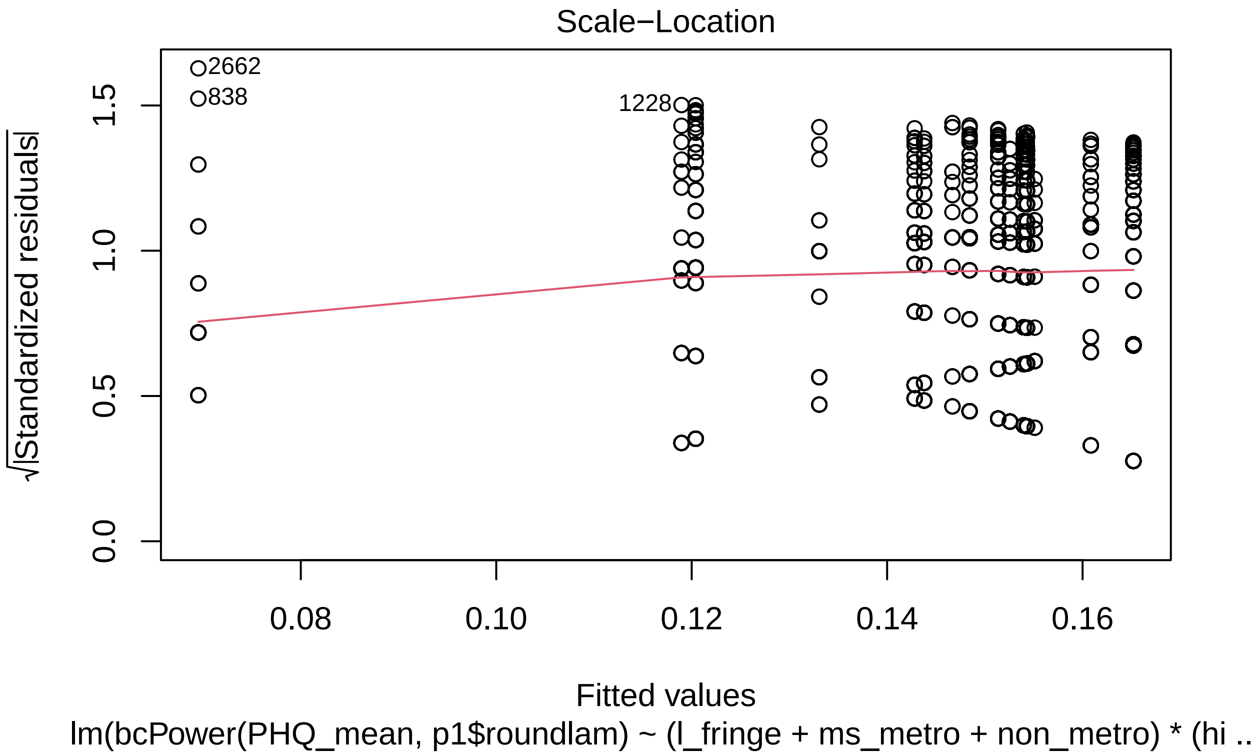
```
summary(m1 <- lm(bcPower(PHQ_mean, p1$roundlam) ~ (l_fringe + ms_metro + non_metro) * (high + bachelor +
```

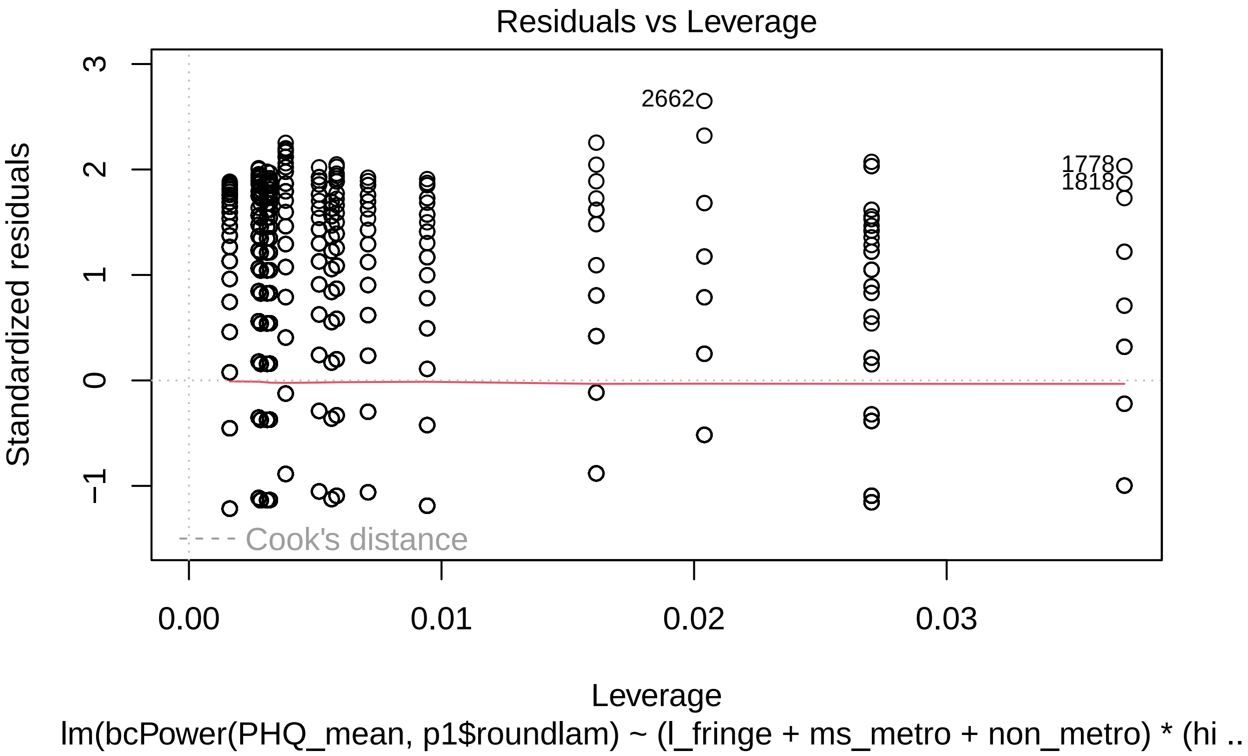
```
##
## Call:
## lm(formula = bcPower(PHQ_mean, p1$roundlam) ~ (l_fringe + ms_metro +
##     non_metro) * (high + bachelor + grad), data = nhis_dummy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16521 -0.14844 -0.01689  0.11468  0.35668
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.06952    0.01943   3.578 0.000352 ***
## l_fringe             0.06354    0.03260   1.949 0.051343 .
## ms_metro             0.04944    0.02600   1.902 0.057290 .
## non_metro            0.07717    0.02962   2.605 0.009228 **
## high                 0.05089    0.02118   2.404 0.016294 *
## bachelor             0.09569    0.02018   4.741 2.22e-06 ***
## grad                 0.08480    0.02074   4.089 4.43e-05 ***
## l_fringe:high       -0.04114    0.03505  -1.174 0.240648
## l_fringe:bachelor   -0.07446    0.03391  -2.196 0.028167 *
## l_fringe:grad       -0.06529    0.03492  -1.869 0.061663 .
## ms_metro:high       -0.01587    0.02839  -0.559 0.576218
## ms_metro:bachelor   -0.06328    0.02751  -2.300 0.021496 *
## ms_metro:grad       -0.05997    0.02932  -2.045 0.040895 *
## non_metro:high      -0.04914    0.03250  -1.512 0.130703
## non_metro:bachelor -0.08153    0.03289  -2.479 0.013237 *
## non_metro:grad      -0.07638    0.03781  -2.020 0.043480 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.136 on 3214 degrees of freedom
## Multiple R-squared:  0.01337,    Adjusted R-squared:  0.008761
## F-statistic: 2.903 on 15 and 3214 DF,  p-value: 0.0001398
```

```
plot(m1)
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
2662
838
12
−2
0
2
Theoretical Quantiles
lm(bcPower(PHQ_mean, p1$roundlam) ~ (l_fringe + ms_metro + non_metro) * (hi ..


Scale−Location
2662
838
1228
1.5
1.0
0.5
0.0
√|Standardized residuals|
0.08
0.10
0.12
0.14
0.16
Fitted values
lm(bcPower(PHQ_mean, p1$roundlam) ~ (l_fringe + ms_metro + non_metro) * (hi ..



```
shapiro.test(m1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m1$residuals
## W = 0.9067, p-value < 2.2e-16
```

Part 4 - Inference

To assess the relationships among education level, regional differences, and mean depression scores, a multiple regression analysis using dummy coded observations for education levels and regional differences was used. In addition, a power analysis was performed to correct the normality violations present. Unfortunately, the power analysis wa not able to fully normalize the residuals. As a result, future analysts should consider more robust procedures.

Based on this analysis, we can conclude that this model predicting changes in mean depression scores based on education levels and regional differences is statistically significant, F(15,3214) = 2.903, p < .001. However, this model may be a relatively poor fit for these data, based on the low Adjusted R^2 value (R^2 = .009). This may be a consequence of the normal, linear, and homoscedasticity violations in this model.

Using this model, we can conclude that living in non-metropolitan regions significantly predicts higher depression scores than living in a metropolitan area (b = 0.077; p < .01). However, living in large fringe or small/medium sized-metropolitan areas versus large central metropolitan areas did not significantly predict changes in mean depression scores.

Interestingly, having a high school education (b = 0.05; p = .016), bachelor's degree (b = 0.10; p < .001), or a graduate to post-graduate degree (b = 0.08; p < .001) predicted higher mean depression scores than having dropped out of high school.

Significant interaction effects emerged such that the role of education moderates the relationship between regional differences and mean depression scores. Specifically, mean depression scores are higher for those who live in a fringe metropolitan area and have a bachelor's degree (b = -0.07; p = .028) compared to those who live in a large central metro area and dropped out of high school. Similarly, living in a medium or small metropolitan area and having a bachelor's degree is associated with higher mean depression scores (b = -0.06; p = .021). In addition, living in a non-metropolitan area and having a bachelor's degree is associated with higher depression scores (b = -0.08; p = .013). Finally, living in a non-metropolitan area and having a graduate or post-graduate degree is associated with higher depression scores, as well (b = -.076; p = .04).

Part 5 - Conclusion

Based on the previous analysis, my hypothesis regarding the association between regional differences and mean depression levels was not supported. It would appear, based on this sample, that living in non-metropolitan areas is associated with higher mean depression ratings compared to large central metropolitan regions. Going back to the results generated in Calhoun's Behavioral Sink experiment, one would expect that living in a more busy, crowded region would predict the presence of psychopathologies. However, our analysis contradicts this argument. This could suggest (as others have suggested previously) that the conclusions of the Behavioral Sink experiment are not generalizable outside of rodent populations, or that psychopathologies are more likely to develop from severe overcrowding conditions, as opposed to the crowding seen in large metropolitan environments. Furthermore, the conclusions derived in this analysis could be a result of the sample characteristics. Clearly, there are far more participants who live in large metropolitan regions versus smaller regions. In the future, a more balanced sample can be employed for analysis.

My hypothesis stating that education levels would predict mean depression scores was incorrect; although the findings were statistically significant, higher education seems to be associated with higher levels of depression. Interestingly, mean depression scores appear to be highest for individuals with a Bachelor's degree (see the **Appendix**) These findings stands in contrast to expectations, but the estimate scores were rather low, indicating that the increases are only marginal. However, these findings may highlight the impact of high-stress jobs associated with higher education. Alternatively, perhaps depression scores just tend to be higher for those in higher education, corresponding with the old adage, "nature shows that with the growth of intelligence comes increased capacity for pain" (Schopenhauer). Of course, further analysis and experimentation are needed to substantiate any causal claims

Finally, my hypothesis claiming that a combination of regional differences and education level would predict mean depression score differences was only partially supported. Based on the results, it would appear that education moderates the relationship between region and mean depression scores, such that higher education is associated with higher depression scores for those not living in metropolitan regions.

Limitations This analysis was limited by a few factors. For one, these data were collected as part of an NHIS study regarding mental health, and thus, these data were not collected for this specific analysis. Furthermore, certain groups were not equivalent, which may have impacted the results of the analysis. Finally, I did not explore the potential influence of extraneous factors that may mediate or moderate the relationships among education, region, and levels of depression. Further research could potentially explore income, race/ethnicity, or other variables as covariates.

References

Bauldry, S. (2015). Variation in the protective effect of higher education against depression. *Society and mental health*, 5(2), 145-161.

Calhoun, J. B. (1962). Population density and social pathology. *Scientific American*, 206(2), 139-149.

Centers for Disease Control and Prevention. (2017, June 1). Data Access - Urban Rural classification scheme for Counties. Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/data_access/urban_rural.htm

Chang-Quan, H., Zheng-Rong, W., Yong-Hong, L., Yi-Zhou, X., & Qing-Xiu, L. (2010). Education and risk for late life depression: a meta-analysis of published literature. *The International Journal of Psychiatry in Medicine*, 40(1), 109-124.

National Center for Health Statistics. (2022). Center for Disease Control and Prevention. https://www.cdc.gov/nchs/nhis/2022nhis.htm

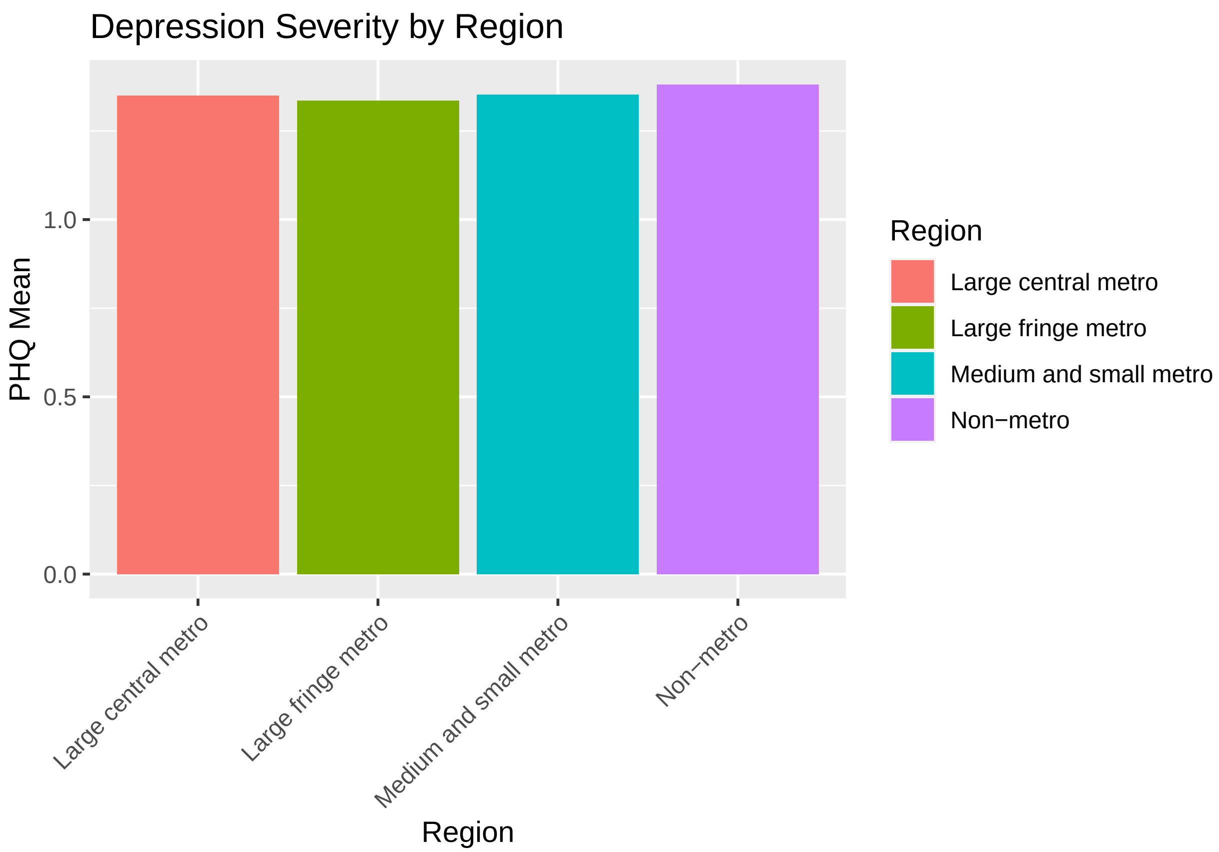
powerTransform function - RDocumentation. (n.d.). https://www.rdocumentation.org/packages/car/versions/2.1-4/topics/powerTransform

Appendix

```
nhis_dep_region <- nhis_dummy %>% group_by(URBRRL) %>% summarize(PHQ_mean = mean(PHQ_mean))

nhis_dep_region <- nhis_dep_region %>%
  mutate(URBRRL =
           dplyr::recode(URBRRL,
                  "1" = "Large central metro",
                  "2" = "Large fringe metro",
                  "3" = "Medium and small metro",
                  "4" = "Non-metro"))

nhis_dep_region <- nhis_dep_region %>% rename(Region = URBRRL)

kable(nhis_dep_region)
```

Region	PHQ_mean
Large central metro	1.350312
Large fringe metro	1.336165
Medium and small metro	1.352765
Non-metro	1.380342

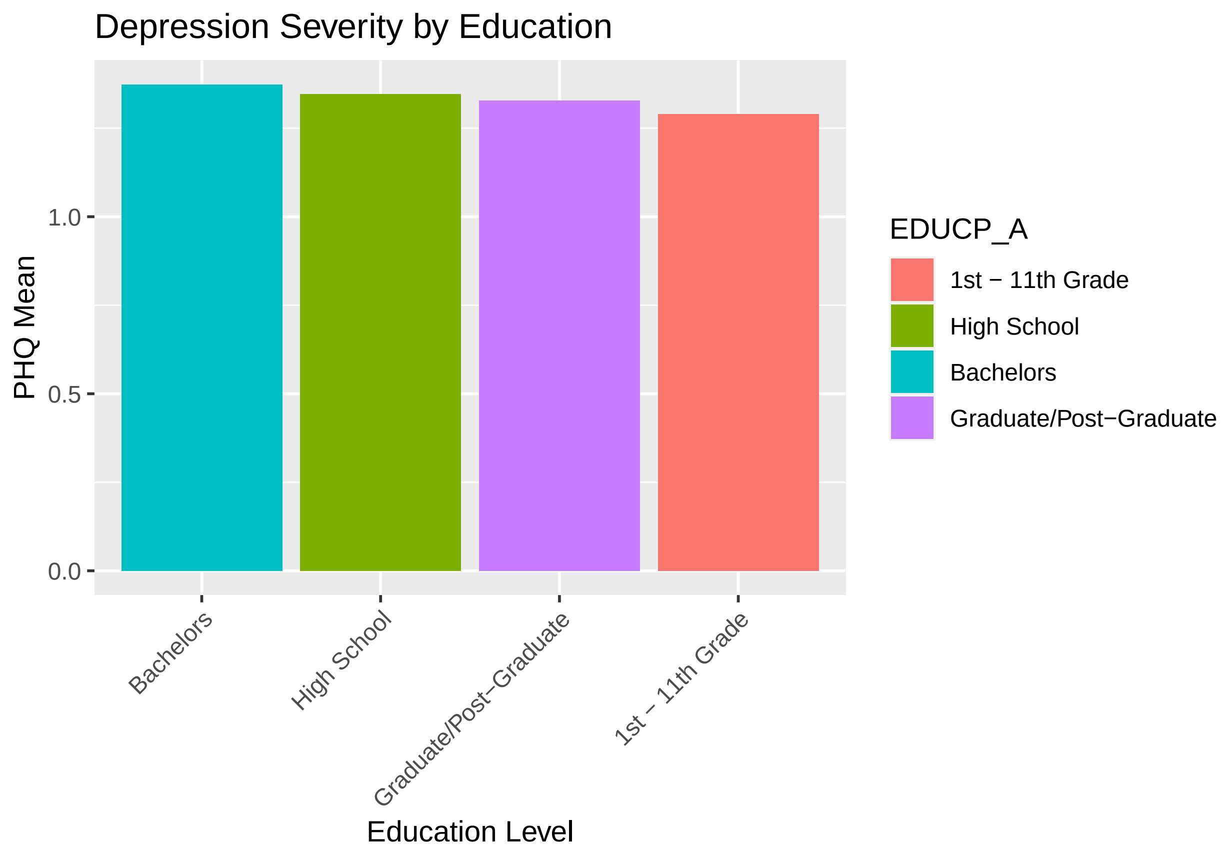
```
ggplot(nhis_dep_region, aes(Region, PHQ_mean, fill = Region)) + geom_bar(stat = "identity")  +  labs(ti
```



```
nhis_dep_edu <- nhis_dummy %>% group_by(EDUCP_A) %>% summarize(PHQ_mean = mean(PHQ_mean))

nhis_dep_edu <- nhis_dep_edu %>% mutate(EDUCP_A = dplyr::recode(EDUCP_A, "1" = "1st - 11th Grade", "2"

kable(nhis_dep_edu)
```

EDUCP_A	PHQ_mean
1st - 11th Grade	1.290000
High School	1.346482
Bachelors	1.373227
Graduate/Post-Graduate	1.328147

```
ggplot(nhis_dep_edu, aes(reorder(EDUCP_A, PHQ_mean, desc), PHQ_mean, fill = EDUCP_A)) + geom_bar(stat =
```


Depression Severity by Education

```
nhis_int <- nhis_dummy %>% group_by(URBRRL, EDUCP_A)  %>% summarize(PHQ_mean = mean(PHQ_mean))
```

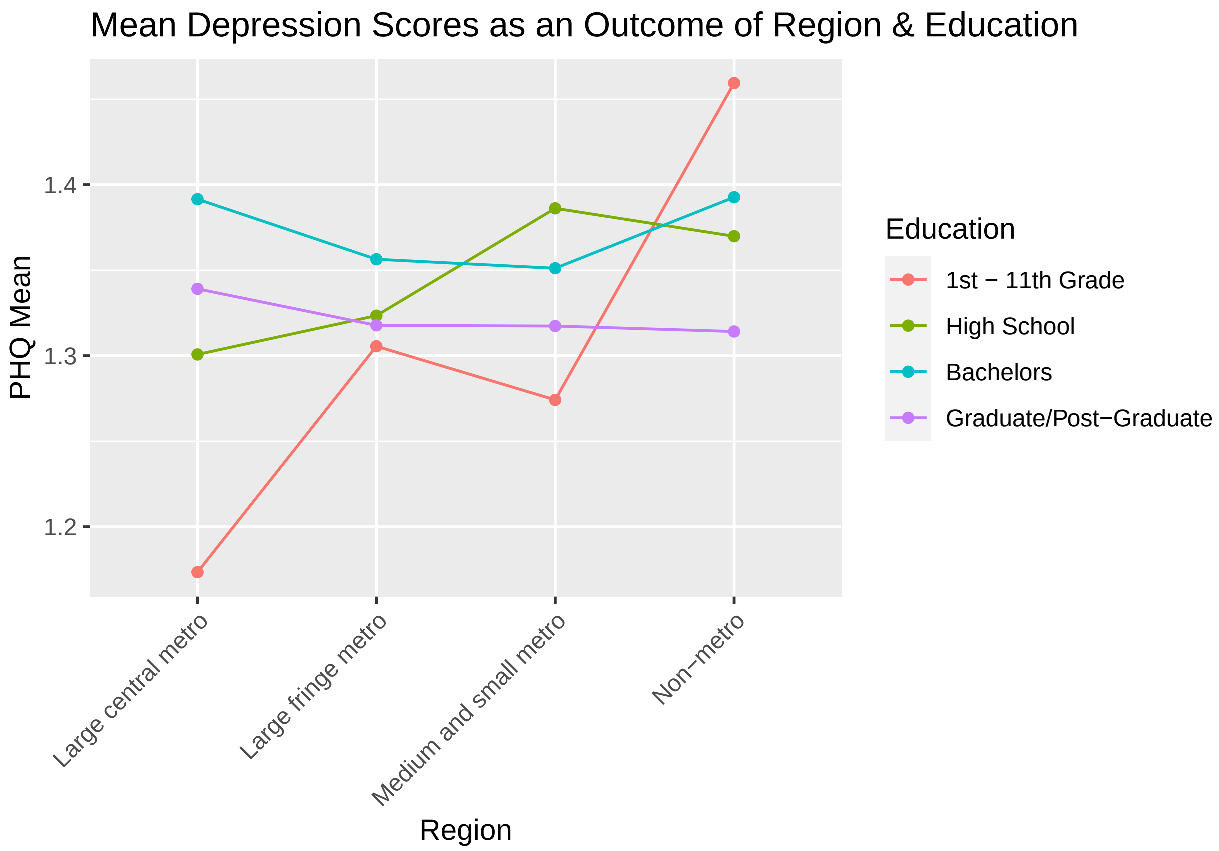
```
## ‘summarise()‘ has grouped output by ’URBRRL’. You can override using the
## ‘.groups‘ argument.
```

```
nhis_int <- nhis_int %>% mutate(URBRRL = dplyr::recode(URBRRL,
                  "1" = "Large central metro",
                  "2" = "Large fringe metro",
                  "3" = "Medium and small metro",
                  "4" = "Non-metro"), EDUCP_A = dplyr::recode(EDUCP_A,
                                                              "1" = "1st - 11th Grade", "2" = "High Sch

nhis_int <- nhis_int %>% rename(Education = EDUCP_A)
nhis_int <- nhis_int %>% rename(Region = URBRRL)

kable(nhis_int)
```

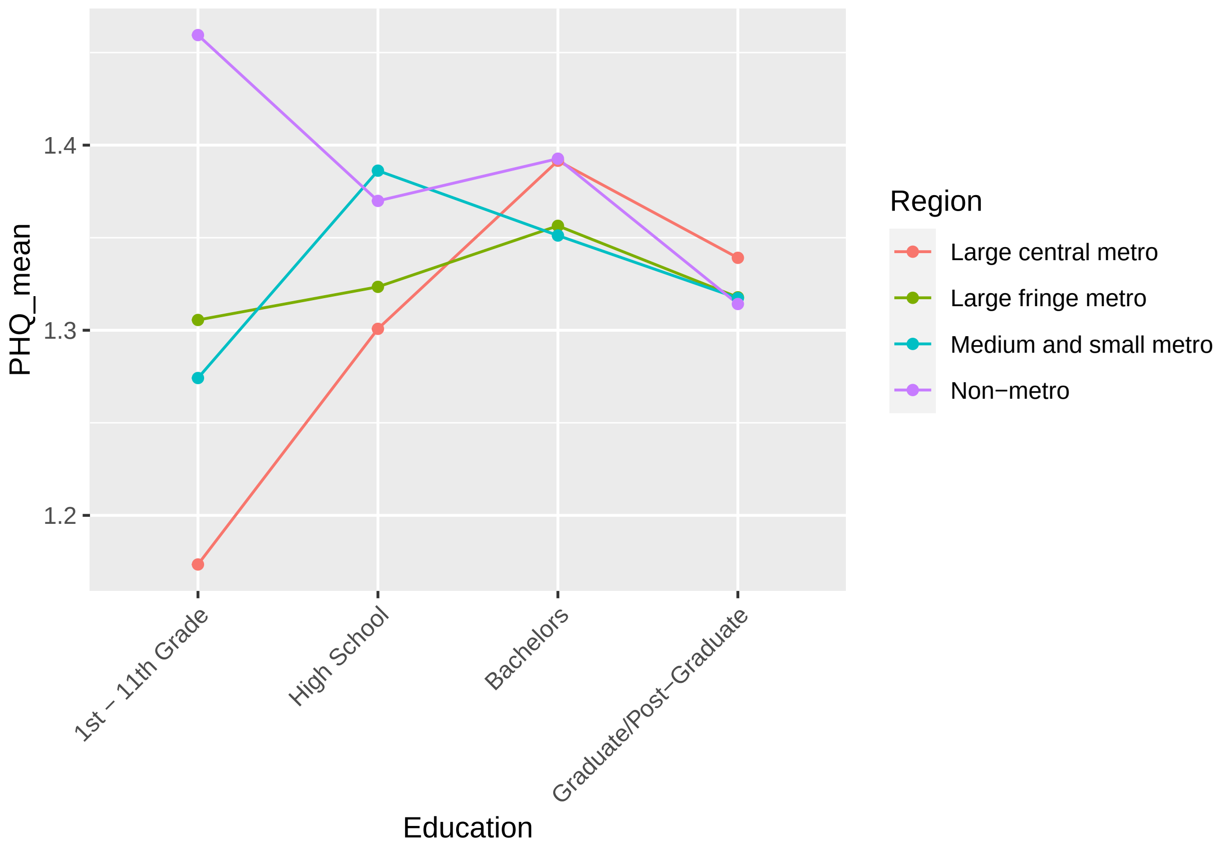
Region	Education	PHQ_mean
Large central metro	1st - 11th Grade	1.173469
Large central metro	High School	1.300766
Large central metro	Bachelors	1.391559
Large central metro	Graduate/Post-Graduate	1.339133
Large fringe metro	1st - 11th Grade	1.305556

Region	Education	PHQ_mean
Large fringe metro	High School	1.323454
Large fringe metro	Bachelors	1.356424
Large fringe metro	Graduate/Post-Graduate	1.317797
Medium and small metro	1st - 11th Grade	1.274193
Medium and small metro	High School	1.386218
Medium and small metro	Bachelors	1.351174
Medium and small metro	Graduate/Post-Graduate	1.317376
Non-metro	1st - 11th Grade	1.459459
Non-metro	High School	1.369883
Non-metro	Bachelors	1.392689
Non-metro	Graduate/Post-Graduate	1.314189

```
nhis_int %>% ggplot(aes(Region, PHQ_mean, group = Education)) + geom_line(aes(color = Education)) + geom
```



```
nhis_int %>% ggplot(aes(Education, PHQ_mean, group = Region)) + geom_line(aes(color = Region)) + geom_p
```


Region
Large central metro
Large fringe metro
Medium and small metro
Non−metro
PHQ_mean
1.4
1.3
1.2
1st − 11th Grade
High School
Bachelors
Graduate/Post−Graduate
Education