Mermaids Part II

Callie Steffen, Vanessa Guenther, Renee Albreight

11/26/2018

#loading packages  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.6  
## v tidyr 0.8.1 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
library(pwr)  
library(kableExtra)  
library(knitr)  
library(plotly)

##   
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

library(ggplot2)  
library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

library(onewaytests)  
library(vcdExtra)

## Loading required package: vcd

## Loading required package: grid

## Loading required package: gnm

##   
## Attaching package: 'vcdExtra'

## The following object is masked from 'package:carData':  
##   
## Burt

## The following object is masked from 'package:plotly':  
##   
## summarise

## The following object is masked from 'package:dplyr':  
##   
## summarise

library(xtable)  
library(ggsignif)  
library(hrbrthemes)

## NOTE: Either Arial Narrow or Roboto Condensed fonts are \*required\* to use these themes.

## Please use hrbrthemes::import\_roboto\_condensed() to install Roboto Condensed and

## if Arial Narrow is not on your system, please see http://bit.ly/arialnarrow

library(babynames)  
library(viridis)

## Loading required package: viridisLite

library(effsize)  
library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

###Reading in csv files  
faculty<-read\_csv("faculty.csv")

## Parsed with column specification:  
## cols(  
## `Faculty Rank` = col\_character(),  
## `Discipline (A = Theoretical, B = Applied)` = col\_character(),  
## `Years Since PhD` = col\_integer(),  
## `Years Faculty Service` = col\_integer(),  
## Sex = col\_character(),  
## `Salary ($)` = col\_integer()  
## )

grad\_enroll<-read\_csv("grad\_enrollment.csv")

## Parsed with column specification:  
## cols(  
## Year = col\_integer(),  
## Total = col\_integer(),  
## Fulltime = col\_integer(),  
## Parttime = col\_integer(),  
## Total\_Males = col\_integer(),  
## Total\_Females = col\_integer(),  
## M\_Fulltime = col\_integer(),  
## M\_Parttime = col\_integer(),  
## F\_Fulltime = col\_integer(),  
## Fpart\_time = col\_integer(),  
## Public = col\_integer(),  
## Total\_Private = col\_integer(),  
## Private\_Nonprofit = col\_character(),  
## Private\_For\_profit = col\_character()  
## )

salary<-read\_csv("mediansalary.csv")

## Warning: Missing column names filled in: 'X3' [3], 'X5' [5]

## Parsed with column specification:  
## cols(  
## `Field of study` = col\_character(),  
## Employment = col\_character(),  
## X3 = col\_character(),  
## `Postdoctoral study` = col\_character(),  
## X5 = col\_character()  
## )

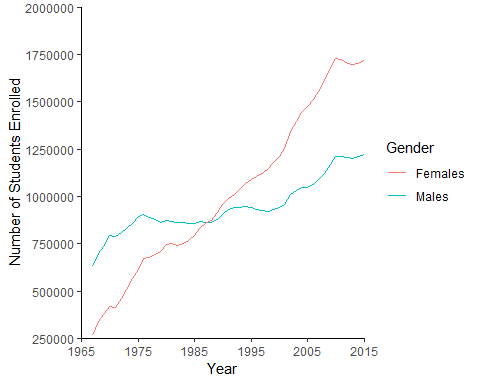
phD<-read\_csv("PhD.csv")

## Warning: Duplicated column names deduplicated: '1985' => '1985\_1' [3],  
## '1990' => '1990\_1' [5], '1995' => '1995\_1' [7], '2000' => '2000\_1' [9],  
## '2005' => '2005\_1' [11], '2015' => '2015\_1' [15]

## Parsed with column specification:  
## cols(  
## FIELD\_SEX = col\_character(),  
## `1985` = col\_character(),  
## `1985\_1` = col\_character(),  
## `1990` = col\_character(),  
## `1990\_1` = col\_character(),  
## `1995` = col\_character(),  
## `1995\_1` = col\_character(),  
## `2000` = col\_character(),  
## `2000\_1` = col\_character(),  
## `2005` = col\_character(),  
## `2005\_1` = col\_character(),  
## `2010` = col\_character(),  
## `201` = col\_character(),  
## `2015` = col\_character(),  
## `2015\_1` = col\_character()  
## )

1. Male and female graduate enrollment (1967 - 2015). Compare trends in total graduate enroll- ment for males and females (including full-time/part-time and private/public universities) in the United States from 1967 - 2015. Describe your results statistically, graphically and in text.

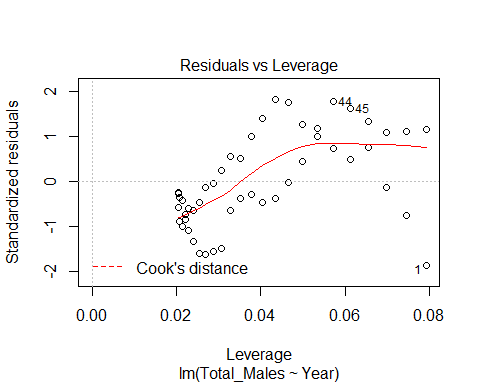
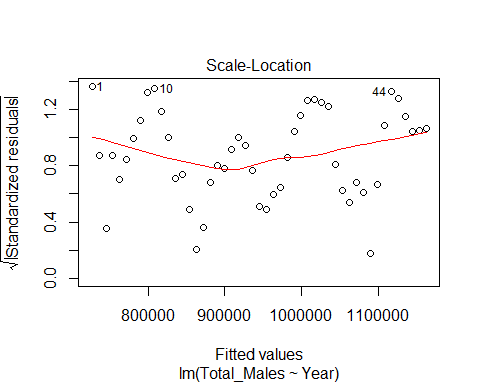
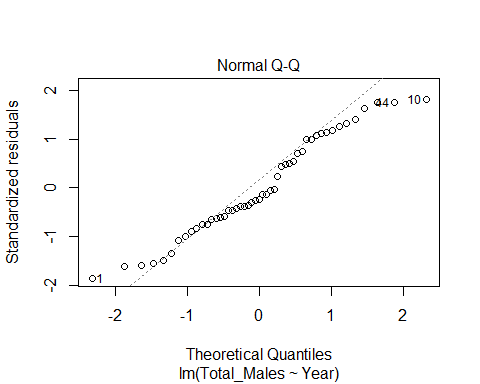
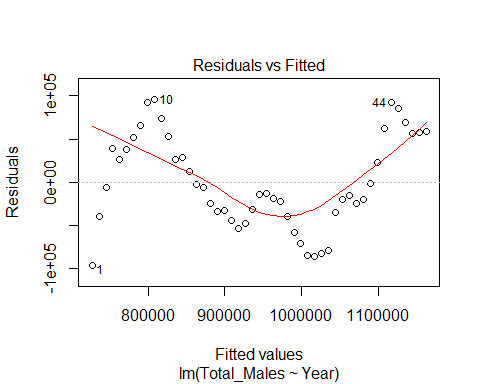
male\_female\_linegraph<- ggplot(grad\_enroll, aes(Year)) +  
 geom\_line(aes(y=Total\_Males, color="Males"))+  
 geom\_line(aes(y=Total\_Females, color="Females"))+  
 labs(colour="Gender") +  
 guides(fill=guide\_legend(title = "Gender"))+  
 scale\_x\_continuous(expand=c(0,0), limits = c(1965,2015), breaks = c(1965,1975,1985,1995,2005,2015))+  
 scale\_y\_continuous(expand = c(0,0), limits=c(250000,2000000),breaks=c(250000,500000,750000,1000000,1250000,1500000,1750000,2000000)) +  
 ylab("Number of Students Enrolled") +  
 theme\_classic()  
   
  
male\_female\_linegraph



##Regression for enrollment on year FOR MALES  
  
regression\_male\_year <- lm(Total\_Males~Year, data=grad\_enroll)  
  
regression\_male\_year

##   
## Call:  
## lm(formula = Total\_Males ~ Year, data = grad\_enroll)  
##   
## Coefficients:  
## (Intercept) Year   
## -17112153 9069

##-17112153+9069(year)  
  
  
plot(regression\_male\_year)



summary(regression\_male\_year)

##   
## Call:  
## lm(formula = Total\_Males ~ Year, data = grad\_enroll)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -96461 -34861 -12841 51876 95766   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -17112153 1087024 -15.74 <2e-16 \*\*\*  
## Year 9069 546 16.61 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 54050 on 47 degrees of freedom  
## Multiple R-squared: 0.8545, Adjusted R-squared: 0.8514   
## F-statistic: 276 on 1 and 47 DF, p-value: < 2.2e-16

##Multiple R-squared: 0.8545, Adjusted R-squared: 0.8514   
##F-statistic: 276 on 1 and 47 DF, p-value: < 2.2e-16  
  
AIC(regression\_male\_year)

## [1] 1210.979

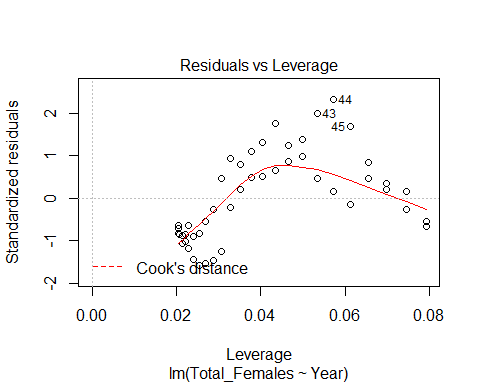
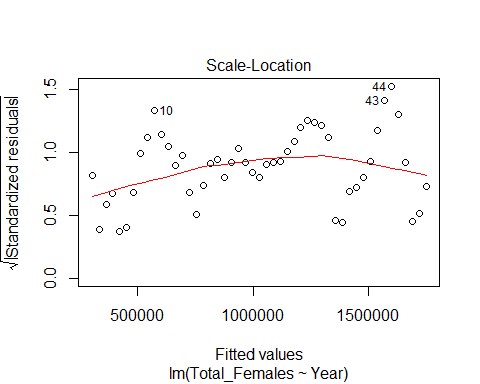
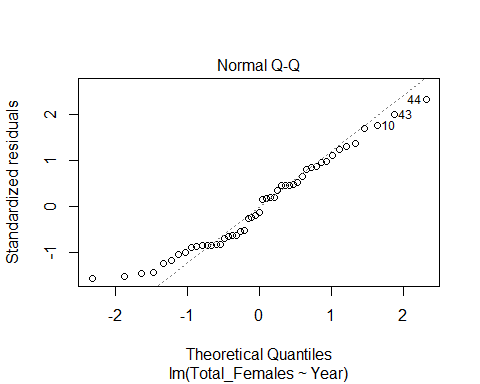
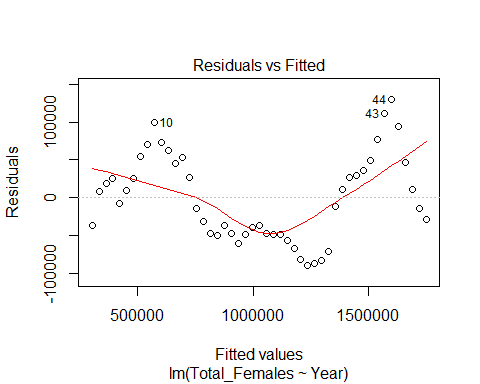
#1210  
  
male\_pr<-cor.test(grad\_enroll$Year, grad\_enroll$Total\_Males)  
  
male\_pr

##   
## Pearson's product-moment correlation  
##   
## data: grad\_enroll$Year and grad\_enroll$Total\_Males  
## t = 16.612, df = 47, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.8690777 0.9568547  
## sample estimates:  
## cor   
## 0.9243741

### Pearson's product-moment correlation  
  
##data: grad\_enroll$Year and grad\_enroll$Total\_Males  
#t = 16.612, df = 47, p-value < 2.2e-16  
#alternative hypothesis: true correlation is not equal to 0  
#95 percent confidence interval:  
# 0.8690777 0.9568547  
#sample estimates:  
#cor   
#0.9243741   
  
  
  
##Regression for enrollment on year FOR FEMALES  
  
regression\_female\_year <- lm(Total\_Females~Year, data = grad\_enroll)  
##(Intercept) -17112153 9069 year  
  
regression\_female\_year

##   
## Call:  
## lm(formula = Total\_Females ~ Year, data = grad\_enroll)  
##   
## Coefficients:  
## (Intercept) Year   
## -58955502 30126

plot(regression\_female\_year)



summary(regression\_female\_year)

##   
## Call:  
## lm(formula = Total\_Females ~ Year, data = grad\_enroll)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -89397 -48101 -7633 45267 129727   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.896e+07 1.161e+06 -50.77 <2e-16 \*\*\*  
## Year 3.013e+04 5.832e+02 51.66 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 57730 on 47 degrees of freedom  
## Multiple R-squared: 0.9827, Adjusted R-squared: 0.9823   
## F-statistic: 2669 on 1 and 47 DF, p-value: < 2.2e-16

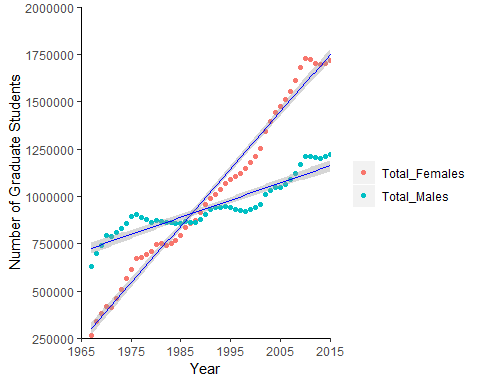
##Multiple R-squared: 0.9827, Adjusted R-squared: 0.9823   
##F-statistic: 2669 on 1 and 47 DF, p-value: < 2.2e-16  
  
  
AIC(regression\_female\_year)

## [1] 1217.443

##1217  
  
female\_pr<-cor.test(grad\_enroll$Year, grad\_enroll$Total\_Females)  
  
female\_pr

##   
## Pearson's product-moment correlation  
##   
## data: grad\_enroll$Year and grad\_enroll$Total\_Females  
## t = 51.659, df = 47, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9845609 0.9951144  
## sample estimates:  
## cor   
## 0.9913086

#Pearson's product-moment correlation  
  
#data: grad\_enroll$Year and grad\_enroll$Total\_Females  
#t = 51.659, df = 47, p-value < 2.2e-16  
#alternative hypothesis: true correlation is not equal to 0  
#95 percent confidence interval:  
# 0.9845609 0.9951144  
#sample estimates:  
 #cor   
#0.9913086   
  
  
#####SMOOTH LINE GRAPH????  
  
smooth <-ggplot(grad\_enroll, aes(x =Year, y = )) +  
 geom\_point(aes(y=Total\_Females, color="Total\_Females"))+  
 geom\_point(aes(y=Total\_Males, color="Total\_Males"))+  
 geom\_smooth(method=lm, se=TRUE, size=.5, color="blue",aes(y=Total\_Females)) +  
 geom\_smooth(method=lm, se=TRUE, size=.5, color="blue",aes(y=Total\_Males))+  
 ylab("Number of Graduate Students") +  
 theme (panel.grid.major = element\_blank(),  
 panel.grid.minor = element\_blank(),panel.background = element\_blank(),axis.line = element\_line(colour = "black")) +  
scale\_x\_continuous(expand = c(0,0), limits = c(1965,2015), breaks = c(1965,1975,1985,1995,2005,2015))+  
 scale\_y\_continuous(expand = c(0,0), limits=c(250000,2000000),breaks=c(250000,500000,750000,1000000,1250000,1500000,1750000,2000000))+  
 scale\_color\_discrete(name = " ")  
  
  
  
smooth

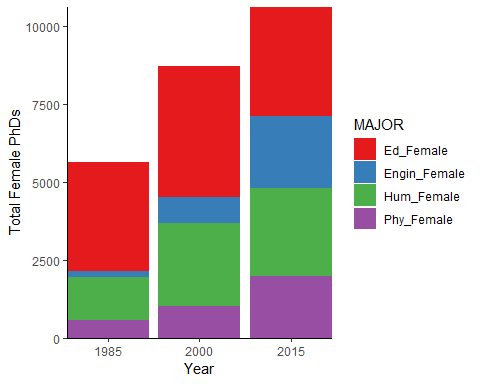


### 2 Females in 4 majors and Phds awarded for 1985 2000 2015

SP<-read\_csv("SPhd.csv")

## Parsed with column specification:  
## cols(  
## Year = col\_integer(),  
## MAJOR = col\_character(),  
## TOTAL\_PHDS = col\_number()  
## )

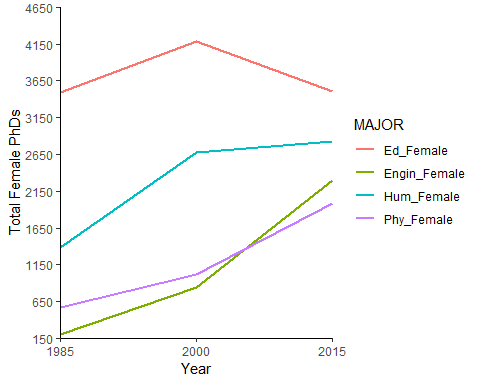
female\_column<- ggplot(SP,aes(x=Year, y=TOTAL\_PHDS))+  
 geom\_col(aes(fill=MAJOR)) +  
 scale\_y\_continuous(expand = c(0,0)) +  
 scale\_x\_continuous(expand = c(0,0), breaks = seq(1985,2015,by=15)) +  
 theme\_classic() +  
 scale\_fill\_brewer(palette = "Set1") +  
 ylab("Total Female PhDs")  
   
female\_column



###PROPORTIONS for FEMALS COLUMN  
  
  
  
linegraph\_phD<- ggplot(SP,aes(x=Year, y=TOTAL\_PHDS, color=MAJOR)) +  
 geom\_line(aes(fill=MAJOR), size=1) +  
 scale\_x\_continuous(expand = c(0,0), limits=c(1985,2015), breaks = c(1985,2000,2015))+  
 scale\_y\_continuous(expand = c(0,0), limits=c(150,4650),breaks=c(150,650,1150,1650,2150,2650,3150,3650,4150,4650)) +  
 ylab("Total Female PhDs") +  
 theme\_classic()

## Warning: Ignoring unknown aesthetics: fill

linegraph\_phD



carmel\_chocolate<-read\_csv("Chocolate .csv")

## Parsed with column specification:  
## cols(  
## `1985` = col\_integer(),  
## `2000` = col\_number(),  
## `2015` = col\_integer()  
## )

##CHI TEST  
  
chi\_female<-chisq.test(carmel\_chocolate)  
  
chi\_female

##   
## Pearson's Chi-squared test  
##   
## data: carmel\_chocolate  
## X-squared = 2073.3, df = 6, p-value < 2.2e-16

##Pearson's Chi-squared test  
  
###data: carmel\_chocolate  
##X-squared = 2073.3, df = 6, p-value < 2.2e-16  
##Null Hypothesis no difference   
## There is a shift in degrees awarded for females throuout the years

3#

#1. The two samples are independent of one another  
# The two populations have equal variance or spread  
 # Read in two seperate data sets  
 employment <- read\_csv("Semployment2.csv")

## Parsed with column specification:  
## cols(  
## Field = col\_character(),  
## Salary = col\_number(),  
## Gender = col\_character()  
## )

postdoc <- read\_csv("Spostdoc2.csv")

## Parsed with column specification:  
## cols(  
## Field = col\_character(),  
## Salary = col\_number(),  
## Gender = col\_character()  
## )

postdoc2 <- read\_csv("Spostdoc.csv")

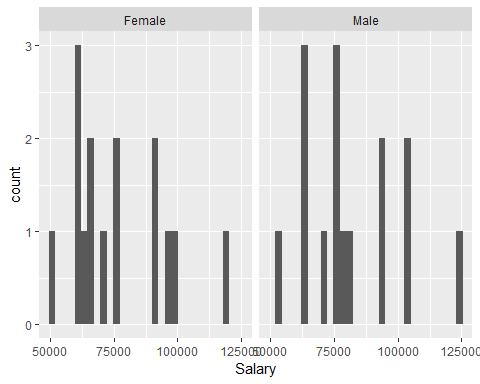
## Parsed with column specification:  
## cols(  
## Field = col\_character(),  
## Male = col\_number(),  
## Female = col\_number()  
## )

employment2 <- read\_csv("SEmployment.csv")

## Parsed with column specification:  
## cols(  
## Field = col\_character(),  
## Male = col\_number(),  
## Female = col\_number()  
## )

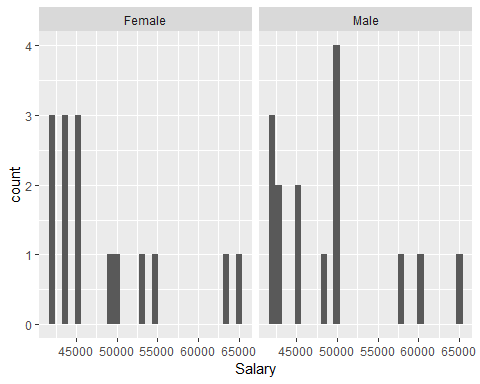
#Explore data Employment   
employment\_hist <- ggplot(employment, aes(x = Salary))+  
 geom\_histogram() +  
 facet\_wrap(~ Gender)  
 employment\_hist

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# not ND   
 #Explore Data Postdoc  
 postdoc\_hist <- ggplot(postdoc, aes(x = Salary))+  
 geom\_histogram() +  
 facet\_wrap(~ Gender)  
 postdoc\_hist

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# not ND  
 #is there a significant difference between male and female post-doc position salaries?  
#Wilcoxon Signed-Rank: Paired  
#he null hypothesis of the Wilcoxon test is usually taken as equal medians  
#If we reject the null, that means we have evidence that one distribution is shifted to the left or right of the other. Since we’re assuming our distributions are equal, rejecting the null means we have evidence that the medians of the two populations differ  
#(n=15)  
 mwu\_postdoc <- postdoc %>%   
 wilcox.test(Salary ~ Gender, paired = TRUE, data = .)

## Warning in wilcox.test.default(x = c(44000, 42000, 43250, 42000, 50000, :  
## cannot compute exact p-value with ties

## Warning in wilcox.test.default(x = c(44000, 42000, 43250, 42000, 50000, :  
## cannot compute exact p-value with zeroes

mwu\_postdoc

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: Salary by Gender  
## V = 16.5, p-value = 0.8884  
## alternative hypothesis: true location shift is not equal to 0

wsr\_postdoc = wilcox.test(postdoc2$Male, postdoc2$Female, paired = TRUE)

## Warning in wilcox.test.default(postdoc2$Male, postdoc2$Female, paired =  
## TRUE): cannot compute exact p-value with ties

## Warning in wilcox.test.default(postdoc2$Male, postdoc2$Female, paired =  
## TRUE): cannot compute exact p-value with zeroes

wsr\_postdoc

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: postdoc2$Male and postdoc2$Female  
## V = 19.5, p-value = 0.8884  
## alternative hypothesis: true location shift is not equal to 0

#No significant difference - distributions are equal  
# what is V!the sum of the ranks in one of both groups.  
  
 #significant different in non-postdoc employment between males and females  
 wsr\_employment = wilcox.test(employment2$Male, employment2$Female, paired = TRUE)

## Warning in wilcox.test.default(employment2$Male, employment2$Female, paired  
## = TRUE): cannot compute exact p-value with ties

## Warning in wilcox.test.default(employment2$Male, employment2$Female, paired  
## = TRUE): cannot compute exact p-value with zeroes

wsr\_employment

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: employment2$Male and employment2$Female  
## V = 101, p-value = 0.002572  
## alternative hypothesis: true location shift is not equal to 0

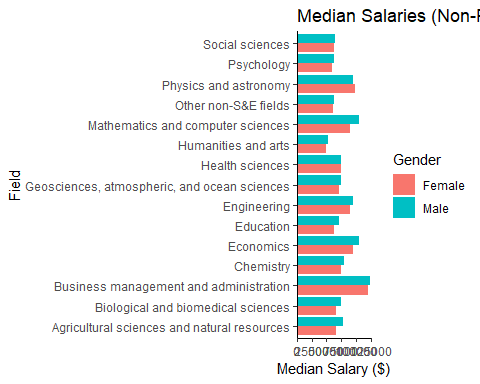
#There is a significant difference  
 #EffectSize  
test <- cliff.delta(employment2$Male, employment2$Female)  
test

##   
## Cliff's Delta  
##   
## delta estimate: 0.2133333 (small)  
## 95 percent confidence interval:  
## inf sup   
## -0.2155378 0.5732121

# small effect size Cliff's Delta ( 0.213333) ? 0 -> groups are totally overlapping

Graph

e\_graph <- ggplot(employment, aes(x = Field, y = Salary ))+  
 geom\_col(position = "dodge", aes(fill = Gender)) +  
 coord\_flip()+  
 scale\_y\_continuous(expand = c(0,0), breaks = seq(0, 125000, 25000), limits = c(0,125000)) +  
 labs(title = "Median Salaries (Non-PostDoc Positions) ", y = "Median Salary ($)")+  
 theme\_classic()  
  
e\_graph



#only use employment grah in   
#Literature ?

# 4 Exploring academic salaries for professors in U.S. colleges.

fac\_salary <- read\_csv("faculty\_salary.csv")

## Parsed with column specification:  
## cols(  
## Rank = col\_character(),  
## Discipline = col\_character(),  
## Years\_Since\_PhD = col\_integer(),  
## Years\_Faculty\_Service = col\_integer(),  
## Sex = col\_character(),  
## Salary = col\_integer()  
## )

#Linear Model 1 - starting with saturated model   
salary\_lm1 <- lm(Salary ~ Sex + Years\_Faculty\_Service + Rank + Years\_Since\_PhD + Discipline, data = fac\_salary)  
#summary(salary\_lm1)  
#plot(salary\_lm1)  
#vif(salary\_lm1) #VIF for Years\_Faculty\_Service = 5.9, VIF for Years\_Since\_PhD = 7.5  
AIC(salary\_lm1)

## [1] 9093.826

#AIC = 9093.826

#linear model 2   
salary\_lm2 <- lm(Salary ~ Sex + Rank + Years\_Since\_PhD + Discipline, data = fac\_salary)  
#summary(salary\_lm2)  
#plot(salary\_lm2)  
#vif(salary\_lm2) #Years since PhD, VIF = 2.06, all others lower  
AIC(salary\_lm2)

## [1] 9097.22

#AIC = 9097.22

#linear model 3   
  
#make prof the reference level  
fac\_salary$Rank<- fct\_relevel(fac\_salary$Rank, "Prof")  
salary\_lm3 <- lm(Salary ~ Sex + Rank + Discipline, data = fac\_salary)  
summary(salary\_lm3)

##   
## Call:  
## lm(formula = Salary ~ Sex + Rank + Discipline, data = fac\_salary)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -66268 -14127 -1566 10813 97718   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 115627 4009 28.841 < 2e-16 \*\*\*  
## SexMale 4492 3860 1.164 0.245   
## RankAssocProf -33680 3177 -10.600 < 2e-16 \*\*\*  
## RankAsstProf -47403 3133 -15.130 < 2e-16 \*\*\*  
## DisciplineB 13709 2295 5.972 5.25e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22640 on 392 degrees of freedom  
## Multiple R-squared: 0.4469, Adjusted R-squared: 0.4412   
## F-statistic: 79.18 on 4 and 392 DF, p-value: < 2.2e-16

#vif(salary\_lm3) #all VIF values lower than 1.04   
AIC(salary\_lm3)

## [1] 9095.454

#AIC = 9095.454  
  
### This model seems to be the best

stargazer(salary\_lm3, type = "html")

Dependent variable:

Salary

SexMale

4,491.801

(3,860.240)

RankAssocProf

-33,679.900\*\*\*

(3,177.338)

RankAsstProf

-47,403.320\*\*\*

(3,133.108)

DisciplineB

13,708.690\*\*\*

(2,295.437)

Constant

115,626.800\*\*\*

(4,009.133)

Observations

397

R2

0.447

Adjusted R2

0.441

Residual Std. Error

22,640.990 (df = 392)

F Statistic

79.180\*\*\* (df = 4; 392)

Note:

*p<0.1;* ***p<0.05;*** p<0.01