Project 1

SDS348 Spring 2021 - 3/15/2021

Sungmo Hong (Sh46959)

Introduction:

Associations between exercise related activities and health-associated factors amongst different counties in Texas

The following two datasets were obtained from www.countyhealthrankings.org and were specifically chosen from the 2020 County Health Rankings.

There were two particular datasets that were chosen.

The first dataset was obtained from the “Ranked Measure Data” and contain the variables “State”, “County”, “% Physically Inactive” (Numeric Variable), and “% With Access to Exercise Opportunities” (Numeric Variable).

The second dataset was obtained from “Additional Measure Data” and contain the variables “County”, “Life Expectancy” (Numeric Variable), and “% Adult with Diabetes” (Numeric Variable).

Both datasets contain the variable “County” which displays all the counties that exist within the state of Texas. There is a variable named “State”, however, this is not particularly useful due to this dataset being specifically from Texas, all the “State” observations will be Texas. The “% Physically Inactive” describes the percentage of adults age 20 and over reporting no leisure-time physical activity. The “% With Access to Exercise Opportunities” refers to the percentage of the popultion with inadequate access to locations for physical activity. Furthermore, the “Life Expectancy” variables provides the average lifespan in years. Lastly, “% Adult with Diabetes” displays the prevelance of diabetes in percentages.

These particular datasets were chosen for two reasons: First, I am pre-health major and so it is no surprise that I have interest in health-associated factors such as diabetes and life-expectancy. Secondly, my passion is fitness and I am part of a clinical exercise physiology lab where we look at the impacts of physical inactivity. Putting the two and two together, I thought it would make for an interesting project to explore.

Based on my experience, I expect there to be several associations between the variables. First, I believe that there should be a pretty strong inverse correlation between physical inactivity and availability of exercise opportunities. Secondly, I expect that inactivity and prevelance of diabetes will also be strongly correlated together. It would be interesting to see if there is an association between physical inactivity and life expectancy, but I presume that there will be.

```
#install necessary libraries
library(readxl)
library(dplyr)
library(tidyverse)
#creates dataset one
dataone <- read_excel("dataset 1.xlsx")
#view dataset
head(dataone)
```

```
## # A tibble: 6 x 4
##   State County   `% Physically Inactive` `% With Access to Exercise Opportuniti~
##   <chr> <chr>                      <dbl>                                   <dbl>
## 1 Texas <NA>                        24.4                                    80.5
## 2 Texas Anderson                    23.1                                    26.4
## 3 Texas Andrews                     26.4                                    93.9
## 4 Texas Angelina                    34.5                                    65.1
## 5 Texas Aransas                     36.3                                    80.6
## 6 Texas Archer                      19.3                                    22.5
```

```
#creates dataset two
datatwo <- read_excel("dataset 2.xlsx")
#view dataset
head(datatwo)
```

```
## # A tibble: 6 x 3
##   County   `Life Expectancy` `% Adults with Diabetes`
##   <chr>                <dbl>                    <dbl>
## 1 <NA>                  79.1                     10
## 2 Anderson              73.4                     10.4
## 3 Andrews               77.5                      6.7
## 4 Angelina              76.1                     14.3
## 5 Aransas               77.9                     18.7
## 6 Archer                78.8                     11.5
```

Tidy:

```
#dataset is tidy, will undo the tidy and go through the process of tidying a dataset

#create untidy dataset using 'pivot_wider()'
untidy <- dataone %>% pivot_wider(names_from = "County", values_from = '% Physically Inactive')
#displays untidy dataset
#dataset is no longer tidy because it is not organized in a way to allow each column to be a variable, each row to be an obs
ervation, with every cell having a singular value
head(untidy)
```

```
## # A tibble: 6 x 257
##   State `% With Access ~  `NA` Anderson Andrews Angelina Aransas Archer
##   <chr>            <dbl> <dbl>    <dbl>   <dbl>    <dbl>   <dbl>  <dbl>
## 1 Texas             80.5  24.4     NA      NA       NA      NA     NA
## 2 Texas             26.4  NA       23.1    NA       NA      NA     NA
## 3 Texas             93.9  NA       NA      26.4     NA      NA     NA
## 4 Texas             65.1  NA       NA      NA       34.5    NA     NA
## 5 Texas             80.6  NA       NA      NA       NA      36.3   NA
## 6 Texas             22.5  NA       NA      NA       NA      NA     19.3
## # ... with 249 more variables: Armstrong <dbl>, Atascosa <dbl>, Austin <dbl>,
## #   Bailey <dbl>, Bandera <dbl>, Bastrop <dbl>, Baylor <dbl>, Bee <dbl>,
## #   Bell <dbl>, Bexar <dbl>, Blanco <dbl>, Borden <dbl>, Bosque <dbl>,
## #   Bowie <dbl>, Brazoria <dbl>, Brazos <dbl>, Brewster <dbl>, Briscoe <dbl>,
## #   Brooks <dbl>, Brown <dbl>, Burleson <dbl>, Burnet <dbl>, Caldwell <dbl>,
## #   Calhoun <dbl>, Callahan <dbl>, Cameron <dbl>, Camp <dbl>, Carson <dbl>,
## #   Cass <dbl>, Castro <dbl>, Chambers <dbl>, Cherokee <dbl>, Childress <dbl>,
## #   Clay <dbl>, Cochran <dbl>, Coke <dbl>, Coleman <dbl>, Collin <dbl>,
## #   Collingsworth <dbl>, Colorado <dbl>, Comal <dbl>, Comanche <dbl>,
## #   Concho <dbl>, Cooke <dbl>, Coryell <dbl>, Cottle <dbl>, Crane <dbl>,
## #   Crockett <dbl>, Crosby <dbl>, Culberson <dbl>, Dallam <dbl>, Dallas <dbl>,
## #   Dawson <dbl>, `Deaf Smith` <dbl>, Delta <dbl>, Denton <dbl>, DeWitt <dbl>,
## #   Dickens <dbl>, Dimmit <dbl>, Donley <dbl>, Duval <dbl>, Eastland <dbl>,
## #   Ector <dbl>, Edwards <dbl>, Ellis <dbl>, `El Paso` <dbl>, Erath <dbl>,
## #   Falls <dbl>, Fannin <dbl>, Fayette <dbl>, Fisher <dbl>, Floyd <dbl>,
## #   Foard <dbl>, `Fort Bend` <dbl>, Franklin <dbl>, Freestone <dbl>,
## #   Frio <dbl>, Gaines <dbl>, Galveston <dbl>, Garza <dbl>, Gillespie <dbl>,
## #   Glasscock <dbl>, Goliad <dbl>, Gonzales <dbl>, Gray <dbl>, Grayson <dbl>,
## #   Gregg <dbl>, Grimes <dbl>, Guadalupe <dbl>, Hale <dbl>, Hall <dbl>,
## #   Hamilton <dbl>, Hansford <dbl>, Hardeman <dbl>, Hardin <dbl>, Harris <dbl>,
## #   Harrison <dbl>, Hartley <dbl>, Haskell <dbl>, Hays <dbl>, ...
```

```
#create tidy dataset using 'pivot_longer()'
#Error result due to existence of "NA" observations. 'drop_na()' function used to remove these "NA" observations
retidy <- untidy%>%pivot_longer(cols = c(3:257),names_to = "County", values_to = '% Physical Inactivity' )%>%drop_na()
#displays tidy dataset
#data is now tidy which allows each column to be a variable, each row to have an observation, with every cell having a singu
lar value
head(retidy)
```

```
## # A tibble: 6 x 4
##   State `% With Access to Exercise Opportunitie~ County   `% Physical Inactivit~
##   <chr>                                   <dbl> <chr>                     <dbl>
## 1 Texas                                    80.5 NA                         24.4
## 2 Texas                                    26.4 Anderson                   23.1
## 3 Texas                                    93.9 Andrews                    26.4
## 4 Texas                                    65.1 Angelina                   34.5
## 5 Texas                                    80.6 Aransas                    36.3
## 6 Texas                                    22.5 Archer                     19.3
```

Join:

```
#'Inner_Join()' function was used to join the two datasets because it would match pairs of observations and drop any rows th
at did not match with a key variable. However, for my particular dataset, it did not matter which join function was used bec
ause both datasets had the same observation county row present and would have resulted in the same resulting merged dataset
 with the same total number of rows. There were no variables that were removed.
JoinData <- dataone%>%inner_join(datatwo, by = "County")

#view joined dataset
head(JoinData)
```

```
## # A tibble: 6 x 6
##   State County `% Physically I~ `% With Access ~ `Life Expectanc~
##   <chr> <chr>             <dbl>            <dbl>            <dbl>
## 1 Texas <NA>               24.4             80.5             79.1
## 2 Texas Ander~             23.1             26.4             73.4
## 3 Texas Andre~             26.4             93.9             77.5
## 4 Texas Angel~             34.5             65.1             76.1
## 5 Texas Arans~             36.3             80.6             77.9
## 6 Texas Archer             19.3             22.5             78.8
## # ... with 1 more variable: `% Adults with Diabetes` <dbl>
```

There were no potential issues to deal with because there were no missing values/incorrectly spelled name, etc. between the two dataset.

Summary Statistics:

filter

```
#Dataset was created using 'Filter()' function.
#NA values in County were filtered
#NA values in Physically Active were filtered
#NA values in With Access to Exercise Opportunities were filtered out
#NA values in life expectancy were filtered out
#NA values in Adults with Diabetes were filtered out
#Prior to filtering, there were 255 observations. After filtering, there were 237 observations. This means that 18 observati
ons were removed.

JoinDataEff<- JoinData%>%filter(!is.na(County))%>%filter(!is.na('% Physically Active'))%>%filter(!is.na('% With Access to Ex
ercise Opportunities'))%>%filter(!is.na(`Life Expectancy`))%>%filter(!is.na('% Adults with Diabetes'))

#View dataset after filtering
glimpse(JoinDataEff)
```

```
## Rows: 237
## Columns: 6
## $ State                                  <chr> "Texas", "Texas", "Texas"...
## $ County                                 <chr> "Anderson", "Andrews", "A...
## $ `% Physically Inactive`                <dbl> 23.1, 26.4, 34.5, 36.3, 1...
## $ `% With Access to Exercise Opportunities` <dbl> 26.389887, 93.865819, 65....
## $ `Life Expectancy`                      <dbl> 73.35175, 77.45956, 76.13...
## $ `% Adults with Diabetes`               <dbl> 10.4, 6.7, 14.3, 18.7, 11...
```

select

```
#The variable "State" provided no real statistical benefit and was removed. It is understood that this data is for counties
 within Texas.
JoinDataEff <- JoinDataEff%>%select(-State)

#View dataset after 'select()' function
head(JoinDataEff)
```

```
## # A tibble: 6 x 5
##   County  `% Physically In~ `% With Access to~ `Life Expectanc~ `% Adults with ~
##   <chr>               <dbl>              <dbl>            <dbl>            <dbl>
## 1 Anders~              23.1              26.4              73.4             10.4
## 2 Andrews              26.4              93.9              77.5              6.7
## 3 Angeli~              34.5              65.1              76.1             14.3
## 4 Aransas              36.3              80.6              77.9             18.7
## 5 Archer               19.3              22.5              78.8             11.5
## 6 Armstr~              23.2               2.52             77.4              9.2
```

mutate

```
#Creation of a categorical variable using 'mutate()' function

#Creates a categorical variable called "Listing_exercise_opportunity" where values above 67% are labeled as "high", values b
etween 33% and 67% are labeled as "med", and values less than 33% are labeled as "low".

NewData <- JoinDataEff%>%mutate(listing_exercise_opportunity = case_when(`% With Access to Exercise Opportunities`> 67 ~ "hi
gh", `% With Access to Exercise Opportunities`<33 ~ "low", `% With Access to Exercise Opportunities` >= 33 & `% With Access
 to Exercise Opportunities`<= 67 ~ "med"))
```

Arrange, group_by, summarize

```
#creates an inactivity dataset
#'select()' function is used to specifically select the "% Physically Inactive" column
#'summarize_if()' function used to specifically numeric variables for 10 different summary statistics
Inactivity<-NewData%>%select(`% Physically Inactive`)%>%summarise_if(is.numeric,funs(mean,median,sd,mad,IQR,var,min,max,n_di
stinct,n()))

#creates an Opportunity dataset
#'select()' function is used to specifically select the "% With Access to Exercise Opportunities" column
#'summarize_if()' function used to specifically numeric variables for 10 different summary statistics
Opportunity<-NewData%>%select(`% With Access to Exercise Opportunities`)%>%summarise_if(is.numeric,funs(mean,median,sd,mad,I
QR,var,min,max,n_distinct,n()))

#creates a Life dataset
#'select()' function is used to specifically select the "Life Expectancy" column
#'summarize_if()' function used to specifically numeric variables for 10 different summary statistics
Life<-NewData%>%select(`Life Expectancy`)%>%summarise_if(is.numeric,funs(mean,median,sd,mad,IQR,var,min,max,n_distinct,n()))

#creates a Diabetes dataset
#'select()' function is used to specifically select the "% Adults with Diabetes" column
#'summarize_if()' function used to specifically numeric variables for 10 different summary statistics
Diabetes<-NewData%>%select(`% Adults with Diabetes`)%>%summarise_if(is.numeric,funs(mean,median,sd,mad,IQR,var,min,max,n_dis
tinct,n()))

#using 'rbind()' function, all of the summarized statistics are put into a new dataset
#used 'rownames_to_column()' function to create a new column named "Summary"
Merge <- rbind(Inactivity, Opportunity, Life, Diabetes)%>%rownames_to_column(var = "Summary")

#Changes row 1 column 1 into "Summary_physical_inactivity"
Merge[1,1]<- "summary_physical_inactivity"

#Changes row 2 column 1 into "summary_access_to_exercise_opportunities"
Merge[2,1]<- "summary_access_to_exercise_opportunities"

#changes row 3 column 1 into "summary_Life_Expectancy"
Merge[3,1]<- "summary_Life_Expectancy"

#changes row 4 column 1 into "summary_Adults_with_Diabetes"
Merge[4,1]<- "summary_Adults_with_Diabetes"

#install "kableExra"
install.packages("kableExtra", repos = "http://cran.us.r-project.org")
```

```
## package 'kableExtra' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\sungm\AppData\Local\Temp\Rtmp2Jes9Y\downloaded_packages
```

```
#pull kableExtra package from library
library(kableExtra)


#use 'kbl()' function to create a table
Merge%>%kbl()%>%kable_classic_2(full_width = F)
```

Summary	mean	median	sd	mad	IQR	var	min	max	n_distinct	n
summary_physical_inactivity	27.51435	27.20000	4.683170	4.596060	6.300000	21.932081	16.600	39.60000	127	237
summary_access_to_exercise_opportunities	58.30829	60.62932	22.603936	23.732992	32.715616	510.937906	0.000	97.56681	237	237
summary_Life_Expectancy	77.45920	77.31532	2.415719	2.239261	2.989091	5.835701	71.862	89.65301	237	237

Summary	mean	median	sd	mad	IQR	var	min	max	n_distinct	n
summary_Adults_with_Diabetes	11.48945	10.70000	4.689143	4.744320	6.700000	21.988066	3.500	29.30000	134	237

```
#creation of Inactivity2 dataset from NewData
#uses 'group_by()' function to group this data by listing_exercise_opportunity
#uses 'select()' function to select column `% With Access to Exercise Opportunities`
#uses 'summarize_if()' function to obtain 10 different summary statistics
#uses 'arrange()' function to arrange mean in descending order.
Inactivity2<-NewData%>%group_by(listing_exercise_opportunity)%>%select(`% Physically Inactive`)%>%summarise_if(is.numeric,fu
ns(mean,median,sd,mad,IQR,var,min,max,n_distinct,n()))%>%arrange(desc(mean))

#creation of Opportunity2 dataset from NewData
#uses 'group_by()' function to group this data by listing_exercise_opportunity
#uses 'select()' function to select column `% With Access to Exercise Opportunities`
#uses 'summarize_if()' function to obtain 10 different summary statistics
#uses 'arrange()' function to arrange mean in descending order.
Opportunity2<-NewData%>%group_by(listing_exercise_opportunity)%>%select(`% With Access to Exercise Opportunities`)%>%summari
se_if(is.numeric,funs(mean,median,sd,mad,IQR,var,min,max,n_distinct,n()))%>%arrange(desc(mean))

#creation of Life2 dataset from NewData
#uses 'group_by()' function to group this data by listing_exercise_opportunity
#uses 'select()' function to select column `Life Expectancy`
#uses 'summarize_if()' function to obtain 10 different summary statistics
#uses 'arrange()' function to arrange mean in descending order.
Life2<-NewData%>%group_by(listing_exercise_opportunity)%>%select(`Life Expectancy`)%>%summarise_if(is.numeric,funs(mean,medi
an,sd,mad,IQR,var,min,max,n_distinct,n()))%>%arrange(desc(mean))

#creation of Diabetes2 dataset from NewData
#uses 'group_by()' function to group this data by listing_exercise_opportunity
#uses 'select()' function to select column `% Adults with Diabetes`
#uses 'summarize_if()' function to obtain 10 different summary statistics
#uses 'arrange()' function to arrange mean in descending order.
Diabetes2<-NewData%>%group_by(listing_exercise_opportunity)%>%select(`% Adults with Diabetes`)%>%summarise_if(is.numeric,fun
s(mean,median,sd,mad,IQR,var,min,max,n_distinct,n()))%>%arrange(desc(mean))

#using 'rbind()' function, all of the summarized statistics are put into a new dataset
#'rownames_to_column()' function used to create new column titled "Summary"
Merge2 <- rbind(Inactivity2, Opportunity2, Life2, Diabetes2)%>%rownames_to_column(var = "Summary")

#Changes row 1 column 1 to "summary_physical_inactivity"
Merge2[1,1]<- "summary_physical_inactivity"

#Changes row 2 through 3 column 1 to ""
Merge2[2:3,1]<-""

#Changes row 4 column 1 to "summary_access_to_exercise_opportunities"
Merge2[4,1]<- "summary_access_to_exercise_opportunities"

#Changes row 5 through 6 column 1 to ""
Merge2[5:6,1]<-""

#Changes row 7 column 1 to "summary_Life_Expectancy"
Merge2[7,1]<- "summary_Life_Expectancy"

#Changes row 8 through 9 column 1 to ""
Merge2[8:9,1]<-""

#Changes row 10 column 1 to "summary_Adults_with_Diabetes"
Merge2[10,1]<- "summary_Adults_with_Diabetes"

#Changes row 11 through 12 column 1 to ""
Merge2[11:12,1]<-""

#create table using 'kbl()' function
Merge2%>%kbl()%>%kable_classic_2(full_width = F)
```

Summary	listing_exercise_opportunity	mean	median	sd	mad	IQR	var	min	max	n_distinct
summary_physical_inactivity	med	28.89076	28.30000	4.849235	5.485620	7.250000	23.515083	19.70000	39.60000	86
	low	26.62759	27.20000	3.848460	3.854760	5.100000	14.810640	19.30000	34.30000	26
	high	25.96292	25.70000	4.161550	3.706500	4.900000	17.318496	16.60000	39.50000	66
summary_access_to_exercise_opportunities	high	80.66542	80.51353	8.374080	10.638786	14.855539	70.125209	67.64706	97.56681	89

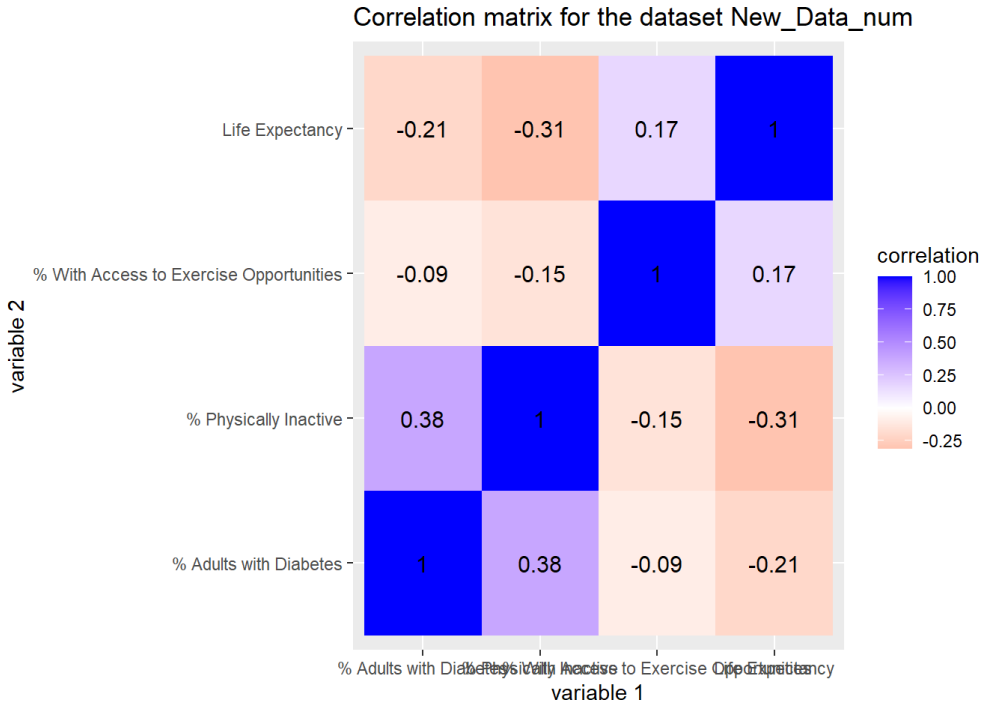
Summary	listing_exercise_opportunity	mean	median	sd	mad	IQR	var	min	max	n_distinct
	med	51.66763	51.85083	9.907872	13.687783	18.426251	98.165919	33.51405	66.89935	119
	low	16.94467	19.30837	11.141357	12.091680	21.917033	124.129838	0.00000	32.41459	29
summary_Life_Expectancy	high	78.09825	78.23461	2.713561	2.541063	3.595904	7.363413	71.86200	89.16425	89
	low	77.47301	77.35534	2.536474	2.537560	3.069493	6.433701	73.35175	85.69699	29
	med	76.97788	76.72424	2.027735	1.512605	2.263951	4.111708	73.34644	89.65301	119
summary_Adults_with_Diabetes	med	12.16050	11.90000	4.810268	5.040840	6.800000	23.138681	3.50000	26.30000	84
	low	11.28276	11.10000	4.319314	5.189100	6.800000	18.656478	4.90000	19.50000	28
	high	10.65955	9.00000	4.548120	3.409980	5.100000	20.685391	3.90000	29.30000	66

Discussion of Overall: This is a summary table of the mean, median, sd, mad, IQR, variance, min, max, n_distinct, and n for four different variables (“Physical Inactivity”, “Access to Exercise Opportunities”, “Life Expectancy”, and “Adults with Diabetes”) prior to being grouped.We can see that Adults with Diabetes has the lowest mean and median while life expectancy has the highest mean and median. There is relatively little variance across the four variables, however, access to exercise opportunities seems to have the highest variability. Immediately, no relationship can be formed just by looking at the table following table.

Discussion of Overall: This is a summary table of the mean, median, sd, mad, IQR, variance, min, max, n_distinct, and n for four different variables (“Physical Inactivity”, “Access to Exercise Opportunities”, “Life Expectancy”, and “Adults with Diabetes”) after being grouped based on “low”, “med”, and “high” level of access of exercise opportunity. Within each grouping, thre is very little variation except for the variable access to exercise opportunities. For the most part, the information provided here is very similar to the first table prior to grouping and does not provide any meaningful relationship right off the bat.

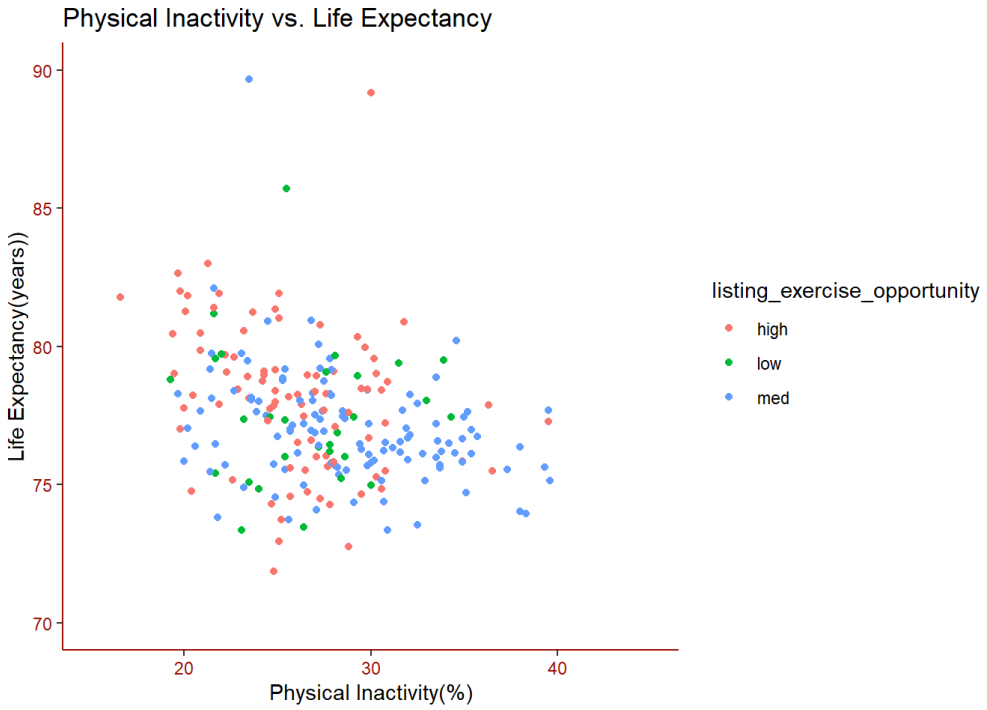
Visualizations

```
#creation of new data set with only numeric variables. This must be done in order to use the 'cor()' functionality.
NewData_num <- NewData%>%select_if(is.numeric)
#use of 'cor()' function for all numeric variables
#save as a data frame
#convert all row names to an explicit variable
#use 'pivot_longer()' function so that all correlations will appear in the same column
#use 'ggplot()' function with aesthetic values
#create a heatmap using 'geom_tile()' function
#change the scaling to make the middle appear neutral
#creation of texts ontop of boxes. Color black and size 4
#creation of title and axis names
#adjust scale and make middle appear neutral
#overlay values on top
#create title and axis
cor(NewData_num, use = "pairwise.complete.obs")%>%as.data.frame%>%rownames_to_column%>%pivot_longer(-1, names_to = "other_va
r", values_to = "correlation")%>%ggplot(aes(rowname, other_var, fill = correlation))+geom_tile()+scale_fill_gradient2(low =
"red", mid = "white", high = "blue")+ geom_text(aes(label = round(correlation,2)), color = "black", size = 4)+labs(title =
"Correlation matrix for the dataset New_Data_num", x = "variable 1", y = "variable 2")
```



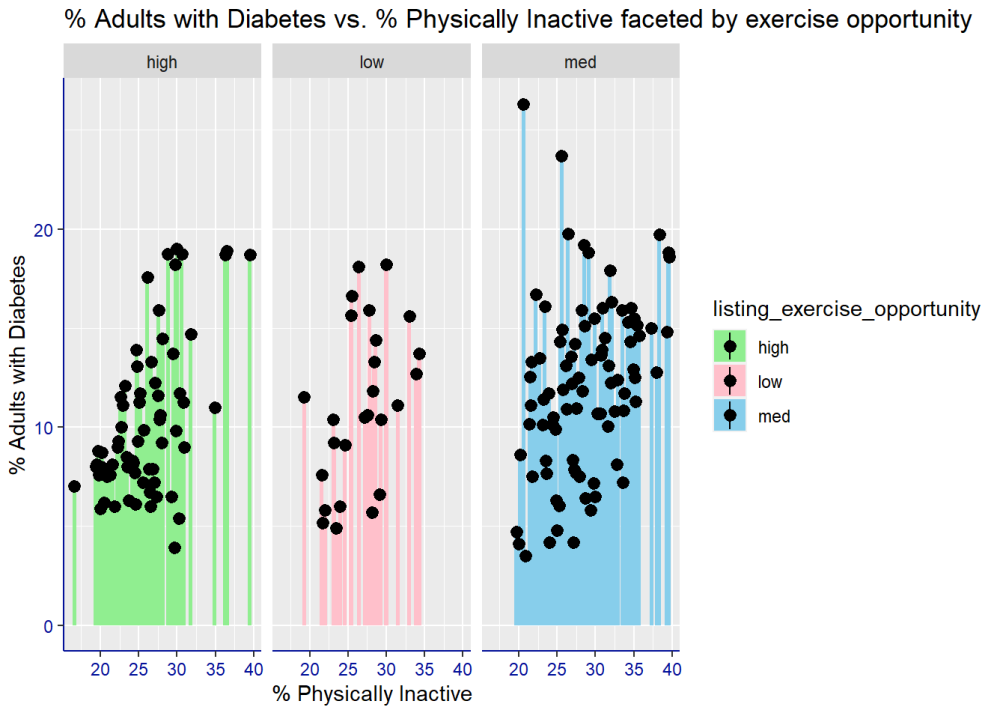
```
#creation of first ggplot
#creation of scatter plot using 'ggplot()' function and 'geom_point()' function
#creation of titles and axis titles
#addition of custom colors to axis and axis values


NewData%>%ggplot()+geom_point(aes(x = `% Physically Inactive`, y = `Life Expectancy`, color = listing_exercise_opportunity))
+labs(title = "Physical Inactivity vs. Life Expectancy", x = "Physical Inactivity(%)", y = "Life Expectancy(years)") + xlim
(15,45)+ylim(70,90)+theme_classic()+theme(axis.line = element_line(color = "#991002"))+theme(axis.text = element_text(color
 = "#991002"))
```




```
#creation of second ggplot
#creation of bar graph using 'ggplot()' and 'geom_bar()'.
#facet graphs based on availability of exercise opportunity from low, med, high
#'stat_summary()' function used to place marker at the mean life expectancy
#create custom colors for axis and axis values


NewData%>%ggplot(aes(x = `% Physically Inactive`, y = `% Adults with Diabetes`, fill = `listing_exercise_opportunity`))+geom
_bar(stat = "summary", width = .5)+stat_summary(fun = mean,color="black",size=.5)+ facet_wrap("listing_exercise_opportunity"
)+scale_fill_manual("listing_exercise_opportunity", values = c(`high` = "light green", `med` = "sky blue", `low` = "pink"))+
theme(axis.line = element_line(color = "#021099"))+theme(axis.text = element_text(color = "#021099"))+labs(title = "% Adults
with Diabetes vs. % Physically Inactive faceted by exercise opportunity")
```



Discussion: Looking at the heat map visualization, the variables "%adults with diabetes" and "% with Access to exercise opportunities" seems to have the lowest correlation between the two. On the other hand, it seems that "% Physically Inactive" and "%Adults with Diabetes" had the strong correlation. However, regardless of how strong the strongest correlation was, there was no particular dataset that had a particularly strong correlation between two datasets.

ggplot1 interpretation:Looking at the visual, there immediately seems to be some correlation between physical inactivity and life expectancy. As physical inactivity increases, the life expectancy decreases. This is pretty expected since inactivity increases the chance of all-cause mortality (information I knew from my lab). There appears to be no correlation between exercise opportunity and life expectancy because the three colors for high, low, med exercise opportunity are randomly distributed as you move up/down the y axis. This observation also applies to the relationship between the three colors for high, low, and med exercise opportunity and physical inactivity.

2nd ggplot interpretation: Looking at the visual, it seems that there is a correlation between increased physical inactivity and the prevalence of adults with diabetes if you exclude the two outliers of diabetes for the "med" group. When faceting based on the availability of exercise opportunity, there does not seem to be a visible difference in regards to the prevalence of adults with diabetes. However, it should be noted that where there is high exercise opportunities available, physical inactivity values tend to be on the slightly lower side compared to low and med exercise opportunity groups.

Dimensionality Reduction: PCA

```
#Prepare data for PCA and run PCA
#selects only numerical variables. Remove categorical variables county and listing_exercise_opportunity
#use 'scale()' function to scale to 0 mean and unit variance (standardization)
pca <- NewData%>%select(c(-County,-listing_exercise_opportunity))%>%scale()%>%prcomp()

#results from PCA
names(pca)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

```
#visualize the pca
pca
```

```
## Standard deviations (1, .., p=4):
## [1] 1.2983557 0.9722974 0.8806030 0.7703561
##
## Rotation (n x k) = (4 x 4):
##                                             PC1        PC2         PC3
## % Physically Inactive                  0.5958952 -0.2243265 -0.07352233
## % With Access to Exercise Opportunities -0.3245080 -0.8709399  0.36755371
## Life Expectancy                        -0.5083198 -0.1393003 -0.80343114
## % Adults with Diabetes                  0.5302966 -0.4144114 -0.46259825
##                                             PC4
## % Physically Inactive                  0.76758128
## % With Access to Exercise Opportunities  0.03259749
## Life Expectancy                         0.27695628
## % Adults with Diabetes                 -0.57710623
```

```
#visualization of the rotated data
head(pca$x)
```

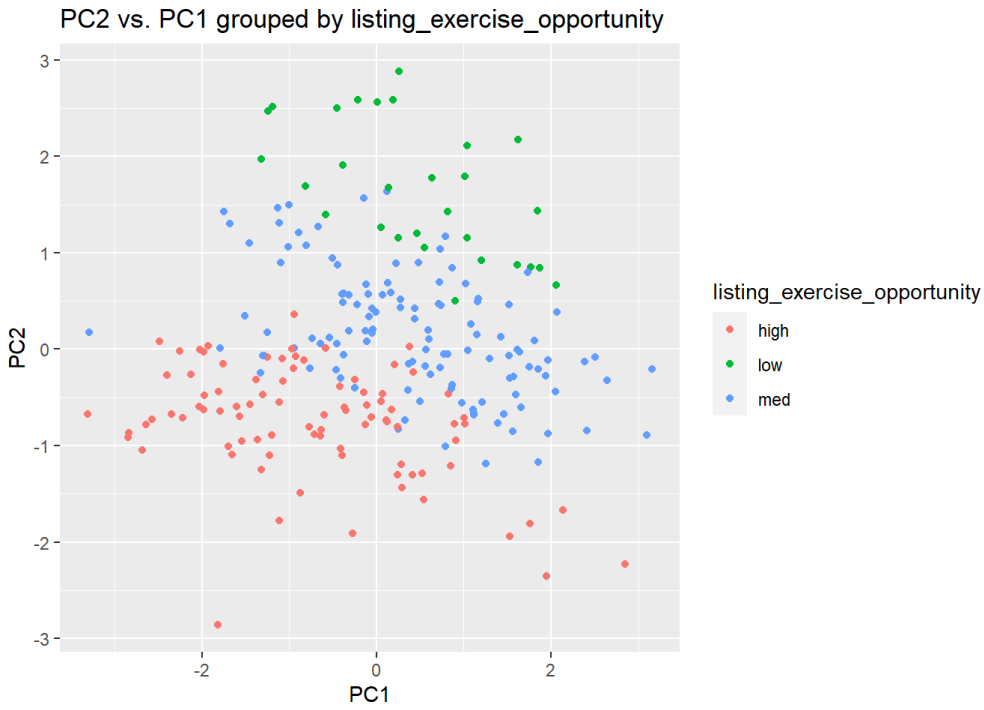
```
##             PC1        PC2        PC3        PC4
## [1,]  0.63762809  1.7744145  1.0238390 -1.1063767
## [2,] -1.19398211 -0.8934145  1.0680533  0.4581279
## [3,]  1.38843460 -0.7691326  0.1659117  0.6565437
## [4,]  1.52543863 -1.9411729 -0.6255869  0.6325754
## [5,] -0.81096015  1.6949144 -0.8981465 -1.2462346
## [6,]  0.01481341  2.5643389 -0.5789358 -0.5177139
```

```
#add information about listing exercise opportunity back into back into the pca data
pca_data <- data.frame(pca$x, listing_exercise_opportunity = NewData$listing_exercise_opportunity)

#view this new combined dataset
head(pca_data)
```

```
##           PC1        PC2        PC3        PC4 listing_exercise_opportunity
## 1  0.63762809  1.7744145  1.0238390 -1.1063767                          low
## 2 -1.19398211 -0.8934145  1.0680533  0.4581279                         high
## 3  1.38843460 -0.7691326  0.1659117  0.6565437                          med
## 4  1.52543863 -1.9411729 -0.6255869  0.6325754                         high
## 5 -0.81096015  1.6949144 -0.8981465 -1.2462346                          low
## 6  0.01481341  2.5643389 -0.5789358 -0.5177139                          low
```

```
# use 'ggplot()' function to plot data according to PC1 and PC2. Title was added using 'labs()' functionality
ggplot(pca_data, aes(x = PC1, y = PC2, color = listing_exercise_opportunity)) + geom_point()+labs(title = "PC2 vs. PC1 group
ed by listing_exercise_opportunity")
```



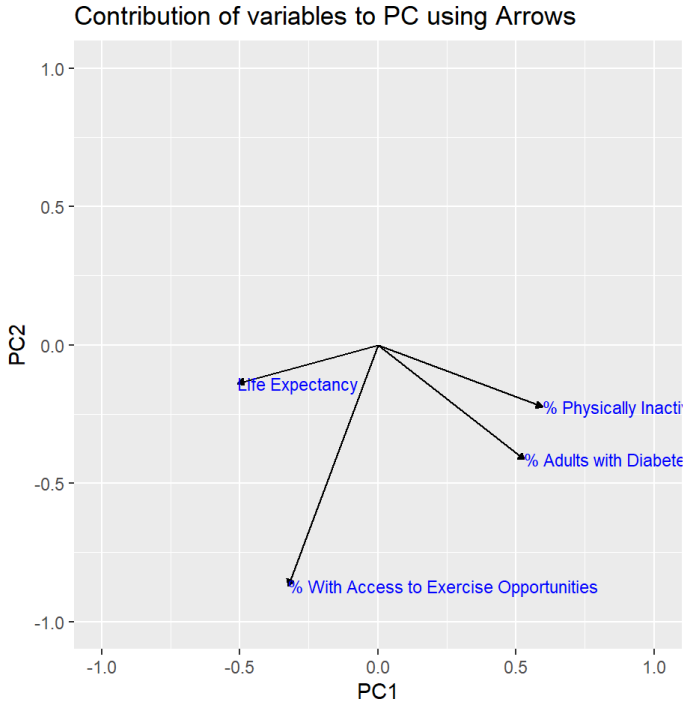

```
#rotation data from pca is viewed
pca$rotation
```

```
##                                           PC1        PC2         PC3
## % Physically Inactive                  0.5958952 -0.2243265 -0.07352233
## % With Access to Exercise Opportunities -0.3245080 -0.8709399  0.36755371
## Life Expectancy                        -0.5083198 -0.1393003 -0.80343114
## % Adults with Diabetes                  0.5302966 -0.4144114 -0.46259825
##                                            PC4
## % Physically Inactive                  0.76758128
## % With Access to Exercise Opportunities  0.03259749
## Life Expectancy                        0.27695628
## % Adults with Diabetes                -0.57710623
```

```
# create a rotation_data dataset
rotation_data <- data.frame(pca$rotation, variable = row.names(pca$rotation))

#establish an arrow style
arrow_style <- arrow(length = unit(0.05, "inches"), type = "closed")

# The contribution of the variables to PCs will be plotted using 'geom_segment()' function.
#geom_segment() for arrows
#geom_text() for labels
ggplot(rotation_data) + geom_segment(aes(xend = PC1, yend = PC2), x = 0, y = 0, arrow = arrow_style) + geom_text(aes(x = PC
1, y = PC2, label = variable), hjust = 0, size = 3, color = "blue") + xlim(-1., 1.) + ylim(-1., 1.) + coord_fixed()+labs(tit
le = "Contribution of variables to PC using Arrows")
```

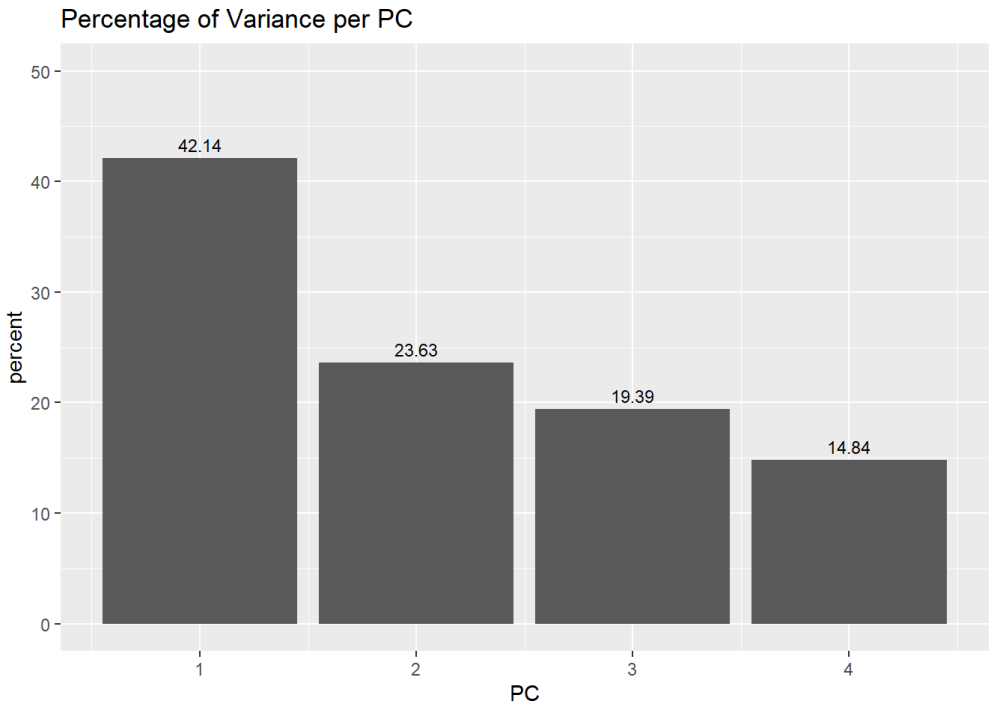



```
# Determine the percentage of variance explained by each component with sdev
percent <- 100* (pca$sdev^2 / sum(pca$sdev^2))
percent
```

```
## [1] 42.14319 23.63406 19.38654 14.83621
```

```
# percent_data dataset created
percent_data <- data.frame(percent = percent, PC = 1:length(percent))

#'ggplot()' function will be made from the percent_data dataset to determine the percentage of variance that can be explaine
d by each principle component
ggplot(percent_data, aes(x = PC, y = percent)) + geom_col() + geom_text(aes(label = round(percent, 2)), size = 3, vjust = -
0.5) + ylim(0, 50)+labs(title = "Percentage of Variance per PC")
```




```
#
```

######Discussion: PC1 and PC2 refer to % Physical Inactivity and %With Access to Exercise Opportunity. Looking at PC1, there is not much variability there among the different group of “high”, “med”, “low”. This would suggest that there is not much of a relationship between the two. However, looking at PC2, we can see that there are distinct “clustering” among “high”, “med”, and “low”. This is expected, considering that PC2 refers to the % With Access to Exercise Opportunity, it can be expected that they would seem “clustered” when you are distinguishing between “high”, “med”, and “low” levels of access to exercise opportunity.

######Discussion: This visual here is showing just how much each variable is contributing to the principle components. We can see that there are four variables, “life expectancy”, “% with access to exercise opportunities”, “% Adults with diabetes”, and “% Physically Inactive”. The arrows are simply a representation of where the principle component is located on a 2D plane. Since this is a rotated dataset, those that deviate most from the horizontal axis represent those components that influence principle component 1 the most while those that deviate most from the vertical axis represent those components that influence priciple component 2 the most.

######Discussion: This visual represents the amount of variation that each principle component contributes in percentages. Based on this, majority of the variation is due to principle component one with principle component two through four having, more or less, similar percentage of variance.

```
##          sysname          release          version         nodename
##        "Windows"         "10 x64"    "build 19042" "DESKTOP-KFAHSG6"
##          machine            login             user    effective_user
##         "x86-64"         "sungm"          "sungm"          "sungm"
```