Project 1: Medical Expenses of Individuals living in Vietnam

Hieu Pham

3/18/2021

Introduction

This project was conducted to observe how medical expenses varied amongst individuals in Vietnam based on their circumstances and status. The datasets that were chosen for this project was acquired through vincentarebundock. The two datasets were generated from a cross sectional study in 1997 and observed a number of 27,765 participants ranging from multiple regions (Cameron, A.C. and P.K. Trivedi). When combined, the resultant dataset included the variables: pharmacy visits, annual medical expenses, age, sex, marriage status, education level, annual number of illnesses/illness days/injuries, number of days inactive, commune population, and insurance status of the individual. Furthermore, I chose to research this topic in order to understand how medical expenses varied in an LDC, such as Vietnam, when compared to a more developed country such as the US. I hypothesize that, similiar to the US, annual medical expenditure and insurance are highly proportional of each other. Those who have a high annual medical expenditure suggests that they can afford healthcare serivces such as insurance, rather than those with a low annual medical expenditure.

Import Datasets

```
## The following datasets were imported into r from excel:
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.6.3
```

```
VietNamI <- read_excel("~/Bioinformatics Data/VietNamI.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
VietNamL <- read_excel("~/Bioinformatics Data/VietNamL.xlsx")
```

```
## New names:
## * `` -> ...1
```

1.) Tidy Datasets

The two datasets acquired were already in a tidy format in that each observation had its own rows and each variable had its own column. For instance, in the “VietnamI” dataset, each observation, representing an individual, was characterized with different factors about the person. These variables were indicated by their own column and included information such as insurance status, commune population, and medical expenditure. Likewise, in the “VietnamL” dataset, each adult, denoted by their respective observation row, was specified by multiple variables, such as their age, years of education, and yearly pharmacy visits. This relationship can be observed in the first few rows by running the codes below:

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```
## -- Attaching packages ---------------------------------------------------------------
------------------------------------------------ tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
## Warning: package 'tibble' was built under R version 3.6.3
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
## Warning: package 'readr' was built under R version 3.6.3
```

```
## Warning: package 'purrr' was built under R version 3.6.3
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
## Warning: package 'stringr' was built under R version 3.6.3
```

```
## Warning: package 'forcats' was built under R version 3.6.3
```

```
## -- Conflicts ------------------------------------------------------------------------
---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
head(VietNamI)
```

```
## # A tibble: 6 x 8
##    ...1 illness injury illdays actdays insurance commune lnhhexp
##   <dbl>   <dbl>  <dbl>   <dbl>   <dbl>     <dbl>   <dbl>   <dbl>
## 1     1       1      0       7       0         0     192    2.73
## 2     2       1      0       4       0         0     167    2.74
## 3     3       0      0       0       0         1      76    2.27
## 4     4       1      0       3       0         1     123    2.39
## 5     5       1      0      10       0         0     148    3.11
## 6     6       0      0       0       0         1      20    3.76
```

```
head(VietNamL)
```

```
## # A tibble: 6 x 7
##    ...1 pharvis lnhhexp   age sex    married  educ
##   <dbl>   <dbl>   <dbl> <dbl> <chr>    <dbl> <dbl>
## 1     1       0    2.73  3.76 male         1     2
## 2     2       0    2.74  2.94 female       0     0
## 3     3       0    2.27  2.56 male         0     4
## 4     4       1    2.39  3.64 female       1     3
## 5     5       1    3.11  3.30 male         1     3
## 6     6       0    3.76  3.37 male         1     9
```

2.) Join/Merge Datasets

```
## Both datasets were combined by linking the individuals via the left_join fuction. The column
 containing individual numbers "...1" was relabed to be "Individual".
library(tidyverse)
VN_1 <- VietNamI
VN_2 <- VietNamL
VN_combined <- VN_1 %>%
  left_join(VN_2, by = "...1") %>%
  rename(Individual = "...1")

## Initial datasets
VietNamI
```

```
## # A tibble: 27,765 x 8
##     ...1 illness injury illdays actdays insurance commune lnhhexp
##    <dbl>   <dbl>  <dbl>   <dbl>   <dbl>     <dbl>   <dbl>   <dbl>
##  1     1       1      0       7       0         0     192    2.73
##  2     2       1      0       4       0         0     167    2.74
##  3     3       0      0       0       0         1      76    2.27
##  4     4       1      0       3       0         1     123    2.39
##  5     5       1      0      10       0         0     148    3.11
##  6     6       0      0       0       0         1      20    3.76
##  7     7       0      0       0       0         1      40    3.16
##  8     8       0      0       0       0         1      57    3.72
##  9     9       2      0       4       0         0      49    2.86
## 10    10       1      0       7       0         0     170    2.62
## # ... with 27,755 more rows
```

```
VietNamL
```

```
## # A tibble: 27,765 x 7
##     ...1 pharvis lnhhexp   age sex    married  educ
##    <dbl>   <dbl>   <dbl> <dbl> <chr>    <dbl> <dbl>
##  1     1       0    2.73  3.76 male         1     2
##  2     2       0    2.74  2.94 female       0     0
##  3     3       0    2.27  2.56 male         0     4
##  4     4       1    2.39  3.64 female       1     3
##  5     5       1    3.11  3.30 male         1     3
##  6     6       0    3.76  3.37 male         1     9
##  7     7       0    3.16  3.66 female       1     2
##  8     8       0    3.72  2.20 male         0     5
##  9     9       2    2.86  3.76 female       1     2
## 10    10       3    2.62  4.23 male         1     0
## # ... with 27,755 more rows
```

```
## Combined dataset
VN_combined
```

```
## # A tibble: 27,765 x 14
##    Individual illness injury illdays actdays insurance commune lnhhexp.x pharvis
##         <dbl>   <dbl>  <dbl>   <dbl>   <dbl>     <dbl>   <dbl>     <dbl>   <dbl>
##  1          1       1      0       7       0         0     192      2.73       0
##  2          2       1      0       4       0         0     167      2.74       0
##  3          3       0      0       0       0         1      76      2.27       0
##  4          4       1      0       3       0         1     123      2.39       1
##  5          5       1      0      10       0         0     148      3.11       1
##  6          6       0      0       0       0         1      20      3.76       0
##  7          7       0      0       0       0         1      40      3.16       0
##  8          8       0      0       0       0         1      57      3.72       0
##  9          9       2      0       4       0         0      49      2.86       2
## 10         10       1      0       7       0         0     170      2.62       3
## # ... with 27,755 more rows, and 5 more variables: lnhhexp.y <dbl>, age <dbl>,
## #   sex <chr>, married <dbl>, educ <dbl>
```

The combined dataset now contains the full list of variables from both individual datsets. No cases were dropped, since the final count of observation was the same as the left dataset, providing 27,765 participants.

3.) Summary Statistics

```
## Install packages for summary statistic tools
install.packages("summarytools", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/hpham/OneDrive/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'summarytools' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\hpham\AppData\Local\Temp\RtmpS4HlAB\downloaded_packages
```

```
library(summarytools)
```

```
## Warning: package 'summarytools' was built under R version 3.6.3
```

```
## Registered S3 method overwritten by 'pryr':
##   method      from
##   print.bytes Rcpp
```

```
## For best results, restart R session and update pander using devtools:: or remotes::install_gi
thub('rapporter/pander')
```

```
##
## Attaching package: 'summarytools'
```

```
## The following object is masked from 'package:tibble':
##
##     view
```

```
## Remove duplicate variable "lnhhexp.y", convert "Insurance" into a categorical variable, filte
r out rows with missing variables, and arrange observation into ascending order.
VN_combined1 <- VN_combined %>%
  select(-lnhhexp.y) %>%
  filter(complete.cases(VN_combined)) %>%
  arrange(Individual) %>%
  mutate(Insurance = case_when(insurance>0 ~ "Yes",
                               insurance<1 ~ "No"))
VN_combined1
```

```
## # A tibble: 27,765 x 14
##    Individual illness injury illdays actdays insurance commune lnhhexp.x pharvis
##         <dbl>   <dbl>  <dbl>   <dbl>   <dbl>     <dbl>   <dbl>     <dbl>   <dbl>
##  1          1       1      0       7       0         0     192      2.73       0
##  2          2       1      0       4       0         0     167      2.74       0
##  3          3       0      0       0       0         1      76      2.27       0
##  4          4       1      0       3       0         1     123      2.39       1
##  5          5       1      0      10       0         0     148      3.11       1
##  6          6       0      0       0       0         1      20      3.76       0
##  7          7       0      0       0       0         1      40      3.16       0
##  8          8       0      0       0       0         1      57      3.72       0
##  9          9       2      0       4       0         0      49      2.86       2
## 10         10       1      0       7       0         0     170      2.62       3
## # ... with 27,755 more rows, and 5 more variables: age <dbl>, sex <chr>,
## #   married <dbl>, educ <dbl>, Insurance <chr>
```

```
##1.) Mean
## The mean of all numerical variables in the dataset.
VN_combined1 %>%
  summarize(mean_illness = mean(illness, na.rm = TRUE), mean_injury = mean(injury, na.rm = TRUE
), mean_illdays = mean(illdays, na.rm = TRUE), mean_actdays = mean(actdays, na.rm = TRUE), mean_
commune = mean(commune, na.rm = TRUE), mean_Inhhexp.x = mean(lnhhexp.x, na.rm = TRUE),  mean_pha
rvis = mean(pharvis, na.rm = TRUE), mean_age = mean(age, na.rm = TRUE), mean_educ = mean(educ, n
a.rm = TRUE))
```

```
## # A tibble: 1 x 9
##   mean_illness mean_injury mean_illdays mean_actdays mean_commune mean_Inhhexp.x
##          <dbl>       <dbl>        <dbl>        <dbl>        <dbl>          <dbl>
## 1        0.622     0.00969         2.80       0.0657         102.           2.60
## # ... with 3 more variables: mean_pharvis <dbl>, mean_age <dbl>,
## #   mean_educ <dbl>
```

```
## The mean of all numerical variables in the dataset grouped by the insurance status of the ind
ividual.
VN_combined1 %>%
  group_by(Insurance) %>%
  summarize(mean_illness = mean(illness, na.rm = TRUE), mean_injury = mean(injury, na.rm = TRUE
), mean_illdays = mean(illdays, na.rm = TRUE), mean_actdays = mean(actdays, na.rm = TRUE), mean_
commune = mean(commune, na.rm = TRUE), mean_Inhhexp.x = mean(lnhhexp.x, na.rm = TRUE),  mean_pha
rvis = mean(pharvis, na.rm = TRUE), mean_age = mean(age, na.rm = TRUE), mean_educ = mean(educ, n
a.rm = TRUE))
```

```
## # A tibble: 2 x 10
##   Insurance mean_illness mean_injury mean_illdays mean_actdays mean_commune
## * <chr>            <dbl>       <dbl>        <dbl>        <dbl>        <dbl>
## 1 No               0.624     0.00933         2.80       0.0677        106.
## 2 Yes              0.612     0.0115          2.85       0.0558         77.7
## # ... with 4 more variables: mean_Inhhexp.x <dbl>, mean_pharvis <dbl>,
## #   mean_age <dbl>, mean_educ <dbl>
```

```
## 2.) Standard Deviation
## The standard deviation of all numerical variables in the dataset.
VN_combined1 %>%
  summarize(sd_illness = sd(illness, na.rm = TRUE), sd_injury = sd(injury, na.rm = TRUE), sd_ill
days = sd(illdays, na.rm = TRUE), sd_actdays = sd(actdays, na.rm = TRUE), sd_commune = sd(commun
e, na.rm = TRUE), sd_Inhhexp.x = sd(lnhhexp.x, na.rm = TRUE),  sd_pharvis = sd(pharvis, na.rm =
TRUE), sd_age = sd(age, na.rm = TRUE), sd_educ = sd(educ, na.rm = TRUE))
```

```
## # A tibble: 1 x 9
##   sd_illness sd_injury sd_illdays sd_actdays sd_commune sd_Inhhexp.x sd_pharvis
##        <dbl>     <dbl>      <dbl>      <dbl>      <dbl>        <dbl>      <dbl>
## 1      0.900    0.0980       5.46       1.12       56.3        0.624       1.31
## # ... with 2 more variables: sd_age <dbl>, sd_educ <dbl>
```

```
## The standard deviation of all numerical variables in the dataset, grouped by the insurance st
atsus of the individual.
VN_combined1 %>%
  group_by(Insurance) %>%
  summarize(sd_illness = sd(illness, na.rm = TRUE), sd_injury = sd(injury, na.rm = TRUE), sd_ill
days = sd(illdays, na.rm = TRUE), sd_actdays = sd(actdays, na.rm = TRUE), sd_commune = sd(commun
e, na.rm = TRUE), sd_Inhhexp.x = sd(lnhhexp.x, na.rm = TRUE),  sd_pharvis = sd(pharvis, na.rm =
TRUE), sd_age = sd(age, na.rm = TRUE), sd_educ = sd(educ, na.rm = TRUE))
```

```
## # A tibble: 2 x 10
##   Insurance sd_illness sd_injury sd_illdays sd_actdays sd_commune sd_Inhhexp.x
## * <chr>          <dbl>     <dbl>      <dbl>      <dbl>      <dbl>        <dbl>
## 1 No             0.899    0.0962       5.42      1.15        55.5        0.612
## 2 Yes            0.903    0.107        5.66      0.904       54.2        0.643
## # ... with 3 more variables: sd_pharvis <dbl>, sd_age <dbl>, sd_educ <dbl>
```

```
## 3.) Variance
## The vatiance of all numerical variables in the dataset.
VN_combined1 %>%
  summarize(var_illness = var(illness, na.rm = TRUE), var_injury = var(injury, na.rm = TRUE), va
r_illdays = var(illdays, na.rm = TRUE), var_actdays = var(actdays, na.rm = TRUE), var_commune =
 var(commune, na.rm = TRUE), var_Inhhexp.x = var(lnhhexp.x, na.rm = TRUE),  var_pharvis = var(ph
arvis, na.rm = TRUE), var_age = var(age, na.rm = TRUE), var_educ = var(educ, na.rm = TRUE))
```

```
## # A tibble: 1 x 9
##   var_illness var_injury var_illdays var_actdays var_commune var_Inhhexp.x
##         <dbl>      <dbl>       <dbl>       <dbl>       <dbl>         <dbl>
## 1       0.809    0.00959        29.8        1.25       3168.         0.390
## # ... with 3 more variables: var_pharvis <dbl>, var_age <dbl>, var_educ <dbl>
```

```
## The variance of all numerical variables in the dataset, grouped by the insurance status of th
e individual.
VN_combined1 %>%
  group_by(Insurance) %>%
  summarize(var_illness = var(illness, na.rm = TRUE), var_injury = var(injury, na.rm = TRUE), va
r_illdays = var(illdays, na.rm = TRUE), var_actdays = var(actdays, na.rm = TRUE), var_commune =
 var(commune, na.rm = TRUE), var_Inhhexp.x = var(lnhhexp.x, na.rm = TRUE),  var_pharvis = var(ph
arvis, na.rm = TRUE), var_age = var(age, na.rm = TRUE), var_educ = var(educ, na.rm = TRUE))
```

```
## # A tibble: 2 x 10
##   Insurance var_illness var_injury var_illdays var_actdays var_commune
## * <chr>           <dbl>      <dbl>       <dbl>       <dbl>       <dbl>
## 1 No              0.808    0.00925        29.4       1.33        3080.
## 2 Yes             0.816    0.0114         32.0       0.817       2940.
## # ... with 4 more variables: var_Inhhexp.x <dbl>, var_pharvis <dbl>,
## #   var_age <dbl>, var_educ <dbl>
```

```
## 4.) N_distinct(number of unique observations)
## The number of unique observations of all numerical variables in the dataset.
VN_combined1 %>%
  summarize(n_illness = n_distinct(illness, na.rm = TRUE), n_injury = n_distinct(injury, na.rm =
TRUE), n_illdays = n_distinct(illdays, na.rm = TRUE), n_actdays = n_distinct(actdays, na.rm = TR
UE), n_commune = n_distinct(commune, na.rm = TRUE), n_Inhhexp.x = n_distinct(lnhhexp.x, na.rm =
TRUE),  n_pharvis = n_distinct(pharvis, na.rm = TRUE), n_age = n_distinct(age, na.rm = TRUE), n_
educ = n_distinct(educ, na.rm = TRUE))
```

```
## # A tibble: 1 x 9
##   n_illness n_injury n_illdays n_actdays n_commune n_Inhhexp.x n_pharvis n_age
##       <int>    <int>     <int>     <int>     <int>       <int>     <int> <int>
## 1         9        2        32        22       194        5735        22    98
## # ... with 1 more variable: n_educ <int>
```

```
## The number of unique observations, grouped by the insurance status of the individual.
VN_combined1 %>%
  group_by(Insurance) %>%
  summarize(n_illness = n_distinct(illness, na.rm = TRUE), n_injury = n_distinct(injury, na.rm =
TRUE), n_illdays = n_distinct(illdays, na.rm = TRUE), n_actdays = n_distinct(actdays, na.rm = TR
UE), n_commune = n_distinct(commune, na.rm = TRUE), n_Inhhexp.x = n_distinct(lnhhexp.x, na.rm =
TRUE),  n_pharvis = n_distinct(pharvis, na.rm = TRUE), n_age = n_distinct(age, na.rm = TRUE), n_
educ = n_distinct(educ, na.rm = TRUE))
```

```
## # A tibble: 2 x 10
##   Insurance n_illness n_injury n_illdays n_actdays n_commune n_Inhhexp.x
## * <chr>         <int>    <int>     <int>     <int>     <int>       <int>
## 1 No                9        2        32        22       194        5512
## 2 Yes               7        2        26        15       194        2350
## # ... with 3 more variables: n_pharvis <int>, n_age <int>, n_educ <int>
```

```
## 5.) Minimum
## The minimum value of each numerical variable in the dataset.
VN_combined1 %>%
  summarize(min_illness = min(illness, na.rm = TRUE), min_injury = min(injury, na.rm = TRUE), mi
n_illdays = min(illdays, na.rm = TRUE), min_actdays = min(actdays, na.rm = TRUE), min_commune =
 min(commune, na.rm = TRUE), min_Inhhexp.x = min(lnhhexp.x, na.rm = TRUE),  min_pharvis = min(ph
arvis, na.rm = TRUE), min_age = min(age, na.rm = TRUE), min_educ = min(educ, na.rm = TRUE))
```

```
## # A tibble: 1 x 9
##   min_illness min_injury min_illdays min_actdays min_commune min_Inhhexp.x
##         <dbl>      <dbl>       <dbl>       <dbl>       <dbl>         <dbl>
## 1           0          0           0           0           1        0.0467
## # ... with 3 more variables: min_pharvis <dbl>, min_age <dbl>, min_educ <dbl>
```

```
## The minimum value of each numerical variable in the dataset, grouped by the insurance status
 of the individual.
VN_combined1 %>%
  group_by(Insurance) %>%
  summarize(min_illness = min(illness, na.rm = TRUE), min_injury = min(injury, na.rm = TRUE), mi
n_illdays = min(illdays, na.rm = TRUE), min_actdays = min(actdays, na.rm = TRUE), min_commune =
 min(commune, na.rm = TRUE), min_Inhhexp.x = min(lnhhexp.x, na.rm = TRUE),  min_pharvis = min(ph
arvis, na.rm = TRUE), min_age = min(age, na.rm = TRUE), min_educ = min(educ, na.rm = TRUE))
```

```
## # A tibble: 2 x 10
##   Insurance min_illness min_injury min_illdays min_actdays min_commune
## * <chr>           <dbl>      <dbl>       <dbl>       <dbl>       <dbl>
## 1 No                  0          0           0           0           1
## 2 Yes                 0          0           0           0           1
## # ... with 4 more variables: min_Inhhexp.x <dbl>, min_pharvis <dbl>,
## #   min_age <dbl>, min_educ <dbl>
```

```
## 6.) Maximum
## The maximum value of each numerical variable in the dataset.
VN_combined1 %>%
  summarize(max_illness = max(illness, na.rm = TRUE), max_injury = max(injury, na.rm = TRUE), ma
x_illdays = max(illdays, na.rm = TRUE), max_actdays = max(actdays, na.rm = TRUE), max_commune =
 max(commune, na.rm = TRUE), max_Inhhexp.x = max(lnhhexp.x, na.rm = TRUE),  max_pharvis = max(ph
arvis, na.rm = TRUE), max_age = max(age, na.rm = TRUE), max_educ = max(educ, na.rm = TRUE))
```

```
## # A tibble: 1 x 9
##   max_illness max_injury max_illdays max_actdays max_commune max_Inhhexp.x
##         <dbl>      <dbl>       <dbl>       <dbl>       <dbl>         <dbl>
## 1           9          1          60          30         194          5.41
## # ... with 3 more variables: max_pharvis <dbl>, max_age <dbl>, max_educ <dbl>
```

```
## The maximum value of each numerical variable in the dataset, grouped by the insurance status
 of the individual.
VN_combined1 %>%
  group_by(Insurance) %>%
  summarize(max_illness = max(illness, na.rm = TRUE), max_injury = max(injury, na.rm = TRUE), ma
x_illdays = max(illdays, na.rm = TRUE), max_actdays = max(actdays, na.rm = TRUE), max_commune =
 max(commune, na.rm = TRUE), max_Inhhexp.x = max(lnhhexp.x, na.rm = TRUE),  max_pharvis = max(ph
arvis, na.rm = TRUE), max_age = max(age, na.rm = TRUE), max_educ = max(educ, na.rm = TRUE))
```

```
## # A tibble: 2 x 10
##   Insurance max_illness max_injury max_illdays max_actdays max_commune
## * <chr>           <dbl>      <dbl>       <dbl>       <dbl>       <dbl>
## 1 No                  9          1          60          30         194
## 2 Yes                 6          1          30          30         194
## # ... with 4 more variables: max_Inhhexp.x <dbl>, max_pharvis <dbl>,
## #   max_age <dbl>, max_educ <dbl>
```

```
## 7.) Median
## The median value of each numerical variable in the dataset.
VN_combined1 %>%
  summarize(med_illness = median(illness, na.rm = TRUE), med_injury = median(injury, na.rm = TRU
E), med_illdays = median(illdays, na.rm = TRUE), med_actdays = median(actdays, na.rm = TRUE), me
d_commune = median(commune, na.rm = TRUE), med_Inhhexp.x = median(lnhhexp.x, na.rm = TRUE),  med
_pharvis = median(pharvis, na.rm = TRUE), med_age = median(age, na.rm = TRUE), med_educ = median
(educ, na.rm = TRUE))
```

```
## # A tibble: 1 x 9
##   med_illness med_injury med_illdays med_actdays med_commune med_Inhhexp.x
##         <dbl>      <dbl>       <dbl>       <dbl>       <dbl>         <dbl>
## 1           0          0           0           0         104          2.53
## # ... with 3 more variables: med_pharvis <dbl>, med_age <dbl>, med_educ <dbl>
```

```
## The median value of each numerical variable in the dataset, grouped by the insurance status o
f the individual.
VN_combined1 %>%
  group_by(Insurance) %>%
  summarize(med_illness = median(illness, na.rm = TRUE), med_injury = median(injury, na.rm = TRU
E), med_illdays = median(illdays, na.rm = TRUE), med_actdays = median(actdays, na.rm = TRUE), me
d_commune = median(commune, na.rm = TRUE), med_Inhhexp.x = median(lnhhexp.x, na.rm = TRUE),  med
_pharvis = median(pharvis, na.rm = TRUE), med_age = median(age, na.rm = TRUE), med_educ = median
(educ, na.rm = TRUE))
```

```
## # A tibble: 2 x 10
##   Insurance med_illness med_injury med_illdays med_actdays med_commune
## * <chr>           <dbl>      <dbl>       <dbl>       <dbl>       <dbl>
## 1 No                  0          0           0           0         111
## 2 Yes                 0          0           0           0          68
## # ... with 4 more variables: med_Inhhexp.x <dbl>, med_pharvis <dbl>,
## #   med_age <dbl>, med_educ <dbl>
```

```
## 8.) Correlation Matrix
## Create vector containing only numerical variables in dataset and construct a  Correlation Mat
rix.
VN_num <- VN_combined1 %>%
  select(illness, injury, illdays, actdays, commune, lnhhexp.x, pharvis, age, educ)%>%
  rename(MExpense = "lnhhexp.x")
VN_Matrix <- VN_combined1 %>%
  select_if(is.numeric)
cor(VN_num, use = "pairwise.complete.obs")
```

```
##              illness        injury     illdays      actdays       commune
## illness   1.00000000  0.0342110387  0.58251926  0.014887487  0.0393017837
## injury    0.03421104  1.0000000000  0.06081290  0.595513059 -0.0006967836
## illdays   0.58251926  0.0608128964  1.00000000  0.081789848  0.0071653205
## actdays   0.01488749  0.5955130586  0.08178985  1.000000000 -0.0097280785
## commune   0.03930178 -0.0006967836  0.00716532 -0.009728078  1.0000000000
## MExpense -0.10070033 -0.0028277489 -0.06495466 -0.009907336 -0.2883442105
## pharvis   0.42627527  0.0482468328  0.35452961  0.045659014  0.0574551630
## age       0.08107781  0.0248544624  0.14656448  0.031475407 -0.0813954238
## educ     -0.04506705 -0.0028733600 -0.02207188 -0.004395189 -0.3294988982
##              MExpense     pharvis         age         educ
## illness  -0.100700333  0.42627527  0.08107781 -0.045067052
## injury   -0.002827749  0.04824683  0.02485446 -0.002873360
## illdays  -0.064954659  0.35452961  0.14656448 -0.022071880
## actdays  -0.009907336  0.04565901  0.03147541 -0.004395189
## commune  -0.288344210  0.05745516 -0.08139542 -0.329498898
## MExpense  1.000000000 -0.03127047  0.06171122  0.255619678
## pharvis  -0.031270466  1.00000000  0.08339587 -0.052768910
## age       0.061711223  0.08339587  1.00000000  0.025132618
## educ      0.255619678 -0.05276891  0.02513262  1.000000000
```

```
## 9. Statistics table for numerical variables
## Install kable package
install.packages("kableExtra", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/hpham/OneDrive/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'kableExtra' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\hpham\AppData\Local\Temp\RtmpS4HlAB\downloaded_packages
```

```
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 3.6.3
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(dplyr)

## Import reformated information from all statistics data and construction of statistics table.
Table1 <- read_excel("~/Bioinformatics Data/Table1.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
Table2 <- read_excel("~/Bioinformatics Data/Table2.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
Table3 <- read_excel("~/Bioinformatics Data/Table3.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
tableA <- Table1 %>%
  rename(Numeric_Variable = "...1")
tableA %>%
  kbl(caption = "Statistics of all Numerical Variables") %>%
  kable_classic(full_width = F, html_font = "Cambria")
```

Statistics of all Numerical Variables

Numeric_Variable	Mean	SD	Variance	Dinstinct Counts	Minimum	Maximum	Median
Illness	0.6219701	0.8995068	0.8091125	9	0.0000000	9.000000	0.000000
Injury	0.0096885	0.0979537	0.0095949	2	0.0000000	1.000000	0.000000
Illdays	2.8040340	5.4582300	29.7922800	32	0.0000000	60.000000	0.000000
Actdays	0.0657302	1.1159390	1.2453190	22	0.0000000	30.000000	0.000000
Commune	101.5266000	56.2833400	3167.8150000	194	1.0000000	194.000000	104.000000
Log of total medical expenditure	2.6026100	0.6244145	0.3898934	5735	0.0467014	5.405502	2.534935
Pharmacy visit	0.5117594	1.3134270	1.7250900	22	0.0000000	30.000000	0.000000
Age	2.9775040	0.9671446	0.9353687	98	0.0000000	4.595120	3.135494
Level of Education	3.3906720	1.9311500	3.7293400	11	0.0000000	11.000000	3.000000

```
## Numerical variables grouped by individuals with insurance
tableB <- Table2 %>%
  rename(Numeric_Variable = "...1")
tableB %>%
  kbl(caption = "Statistics of all Numerical Variables of Individuals with Insurance") %>%
  kable_classic(full_width = F, html_font = "Cambria")
```

Statistics of all Numerical Variables of Individuals with Insurance

Numeric_Variable	Mean	SD	Variance	Dinstinct Counts	Minimum	Maximum	Median
Illness	0.6120	0.903	8.16e-01	7	0.000	6.00	0.00
Injury	0.0121	0.107	1.14e-02	2	0.000	1.00	0.00
Illdays	2.8500	5.660	3.20e+01	26	0.000	30.00	0.00
Actdays	0.0562	0.904	8.17e-01	15	0.000	30.00	0.00
Commune	77.7000	54.200	2.94e+03	194	1.000	194.00	68.00
Log of total medical expenditure	2.8200	0.643	4.13e-01	2350	0.663	5.23	2.76
Pharmacy visit	0.4030	1.030	1.06e+00	13	0.000	20.00	0.00
Age	3.1300	0.787	6.19e-01	88	0.000	4.52	3.14
Level of Education	4.4800	2.290	5.24e+00	11	0.000	11.00	4.00

```
## Numerical variables grouped by individuals without insurance
tableC <- Table3 %>%
  rename(Numeric_Variable = "...1")
tableC %>%
  kbl(caption = "Statistics of all Numerical Variables of Individuals without Insurance") %>%
  kable_classic(full_width = F, html_font = "Cambria")
```

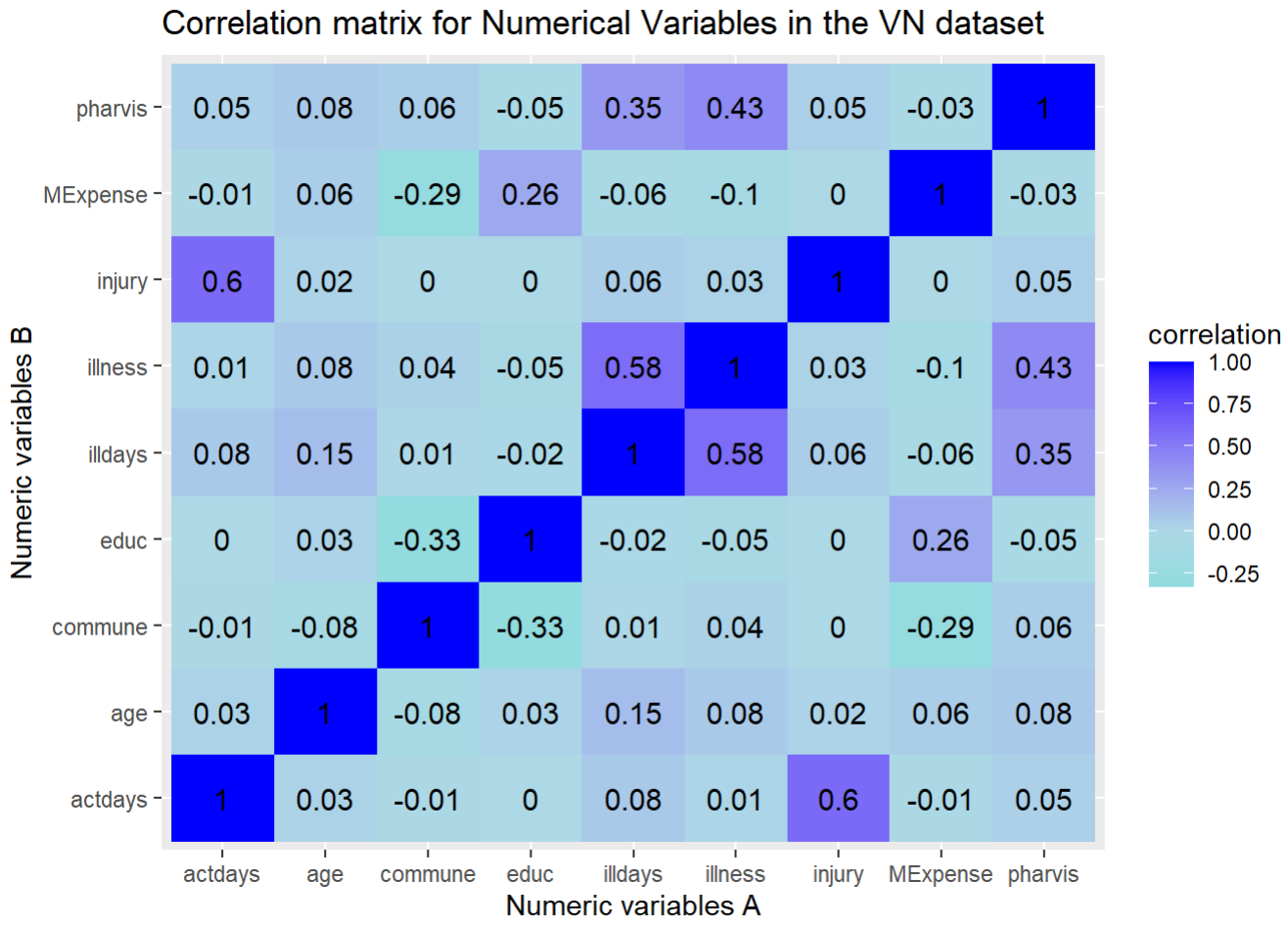
Statistics of all Numerical Variables of Individuals without Insurance

Numeric_Variable	Mean	SD	Variance	Dinstinct Counts	Minimum	Maximum	Median
Illness	6.24e-01	0.8987781	0.8078020	9	0.0000000	9.000000	0.000000
Injury	9.33e-03	0.0961573	0.0092462	2	0.0000000	1.000000	0.000000
Illdays	2.80e+00	5.4191070	29.3667200	32	0.0000000	60.000000	0.000000
Actdays	6.77e-02	1.1526236	1.3285412	22	0.0000000	30.000000	0.000000
Commune	1.06e+02	55.4981500	3080.0440000	194	1.0000000	194.000000	111.000000
Log of total medical expenditure	2.56e+00	0.6121768	0.3747604	5512	0.0467014	5.405502	2.494259
Pharmacy visit	5.33e-01	1.3608360	1.8518740	22	0.0000000	30.000000	0.000000
Age	2.95e+00	0.9955991	0.9912176	98	0.0000000	4.595120	3.135494
Level of Education	3.18e+00	1.7782290	3.1620970	11	0.0000000	11.000000	3.000000

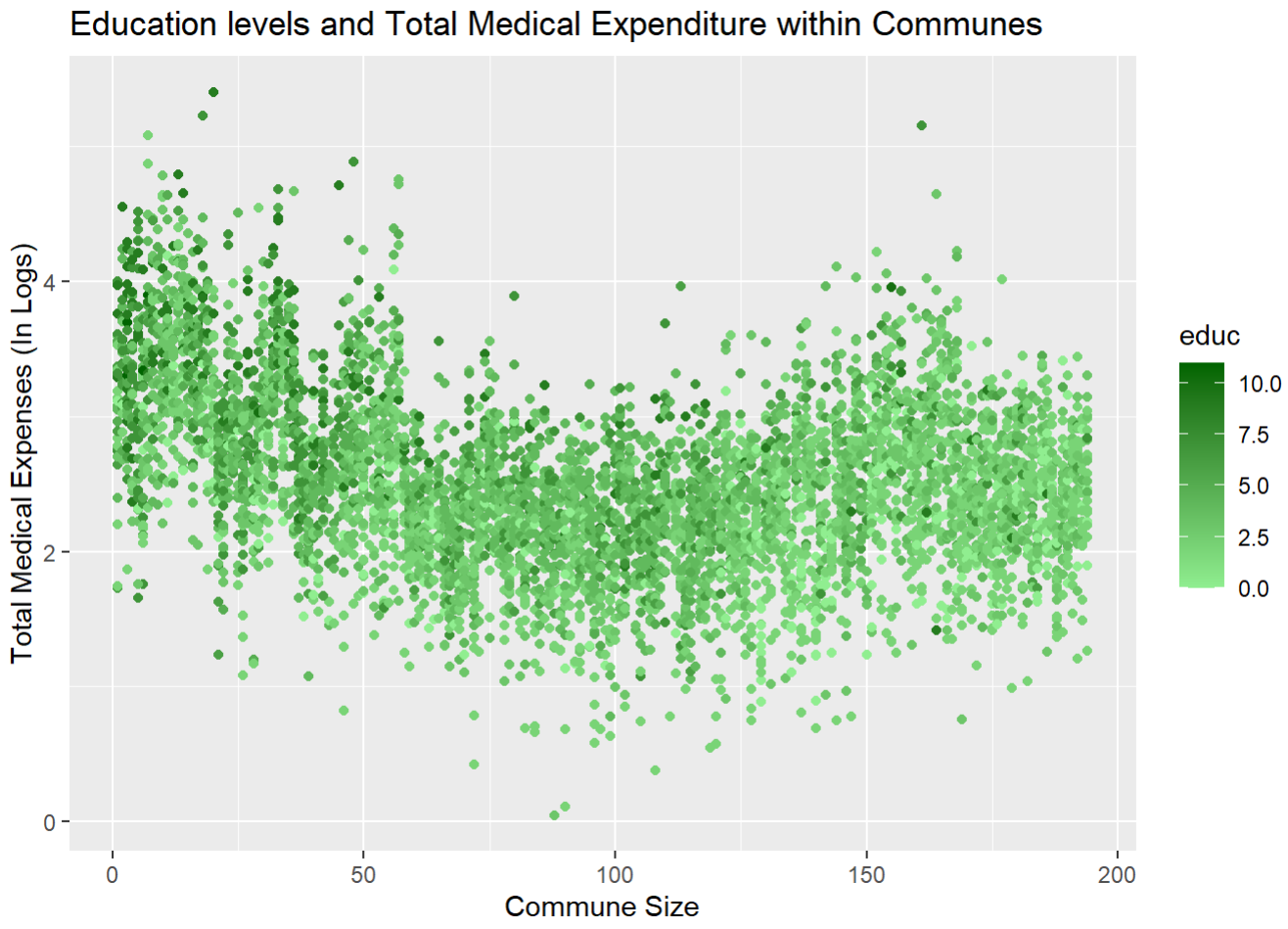
Fourteen different summary statistics were generated using stats functions (mean, median, min, max, N_distinct, variance, and standard deviation) from a combination of all numeric variables and when categorized by insurance status. As expected, the average total medical expenditure was higher in individuals with insurance than in individual without insurance. Moreover, the average level of education was higher in those with insurance, with a higher medical expenditure, than when compared to those without insurance, and with a much lower medical expenditure. However, this relationship only accounts for an r value of 0.25 (education v. MExpense) in the constructed correlation matrix. The numeric variables “Injury” and “Inactive days” displays the highest correlation as indicated by their r value of 0.595.

4.) Visualizations

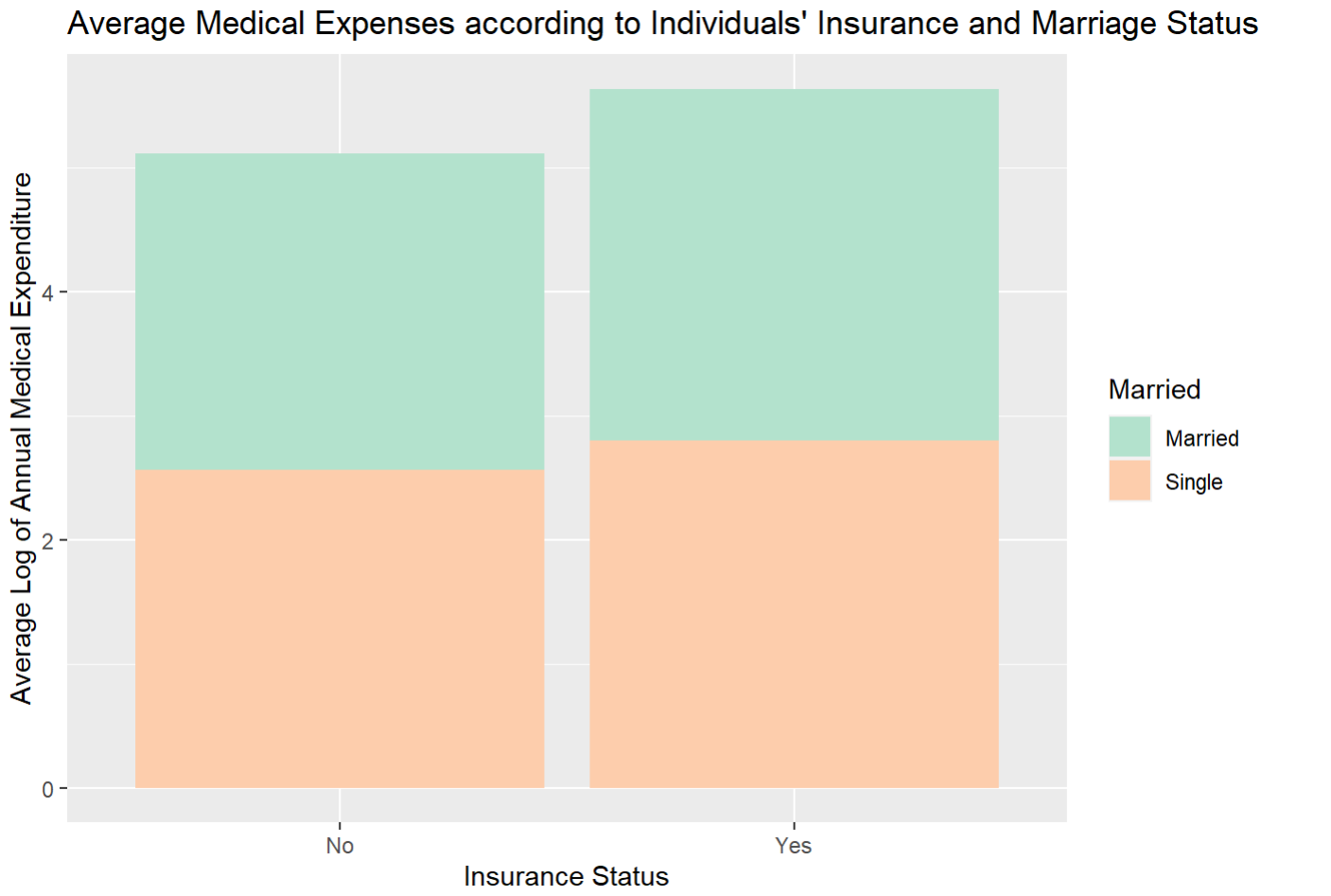
```
##  1.) Heat Map
## This function was used to construct a heat map based on the correlation matrix generated in t
he previous section. The colors were modified to turquoise,lightblue, and blue in order to show
 correlation strength and graph/axes titles were added to complete the heat map.
cor(VN_num, use = "pairwise.complete.obs") %>%
  as.data.frame %>%
  rownames_to_column %>%
  pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>%
  ggplot(aes(rowname, other_var, fill=correlation)) +
  geom_tile() +
  scale_fill_gradient2(low="turquoise",mid="lightblue",high="blue") +
  geom_text(aes(label = round(correlation,2)), color = "black", size = 4) +
  labs(title = "Correlation matrix for Numerical Variables in the VN dataset", x = "Numeric vari
ables A", y = "Numeric variables B")
```



```
## 2.) Scatter plot
## The variable "lnhhexp.x" was renamed in the dataset to be the log total medical expenditure o
f observations.
library(ggplot2)
VN_SP <- VN_combined1%>%
  rename(Log_of_Total_Medical_Expenditure = "lnhhexp.x")
## A scatterplot was graphed to indicate the education levels and total medical expenses of indi
viduals within communes.
ggplot(VN_SP, aes(commune,Log_of_Total_Medical_Expenditure)) +
  geom_point(aes(color=educ)) +
  ggtitle("Education levels and Total Medical Expenditure within Communes") +
  scale_color_gradient(low="lightgreen",high="darkgreen") +
  xlab("Commune Size") +
  ylab("Total Medical Expenses (In Logs)")
```



```
## 3.) Bar plot
## Mutate variable in order to assign marriage status in place of numeric 0 and 1.
VN_BP <- VN_combined1 %>%
  mutate(Married = case_when(married>0 ~ "Married",
                             married<1 ~ "Single"))
## Bar plot displaying the number of married individuals and whether they have insurance. Variab
les include marriage status, insurance status, and the counts of individuals.
ggplot(VN_BP, aes(x = Insurance, y = lnhhexp.x))+
  geom_bar(stat='summary', fun='mean', aes(fill=Married)) +
  scale_fill_brewer(palette = "Pastel2") +
  ggtitle("Average Medical Expenses according to Individuals' Insurance and Marriage Status") +
  xlab("Insurance Status") +
  ylab("Average Log of Annual Medical Expenditure")
```



The heatmap helps to visualize the correlation strength between the multiple numeric variables in this data. As supported by the correlation matrix, injury and inactive days share the highest correlation value of 0.6, and is colored a deep purple to indicate its strength. Quadrants colored light blue are generally around the 0 correlation range, and suggest no relationship between each other.

The scatterplot contains three distinct variables: commune size, annual medical expense, and education level. According to the graph, there is a mild negative relationship between medical expense and commune size, suggesting that smaller commune size invokes higher medical expenses. Moreover, education levels, as indicated by the color spectrum, is lower in larger commune sizes with lower medical expenses.

The barplot contains three unique variables: Insurance status, annual medical expenditure, and the mean of marriage status. According to the graph, the number of individuals with insurance seems to have a higher annual medical expenditure average than the former. Moroever, the mean of those who are married seems to be equal to those who are single in either cases. This may suggest that marriage status does not share a strong correlation to insurance status and medical expenses.

5.) PCA Test

```
## Conducting PCA
## 1.) Data preparation and PCA test
## The dataset was modified to only contain numeric variables which were then scaled. Moreover,
 the dataset was reaffirmed to be "tidy" (a row to every observation and a column to every varia
ble) prior to conducting PCA.
VN_PCA <- VN_combined1 %>%
  select(illness, injury, illdays, actdays, commune,lnhhexp.x,pharvis, age, educ) %>%
  scale() %>%
  as.data.frame %>%
## PCA is conducted with the function "prcomp()"
  prcomp()
VN_PCA
```

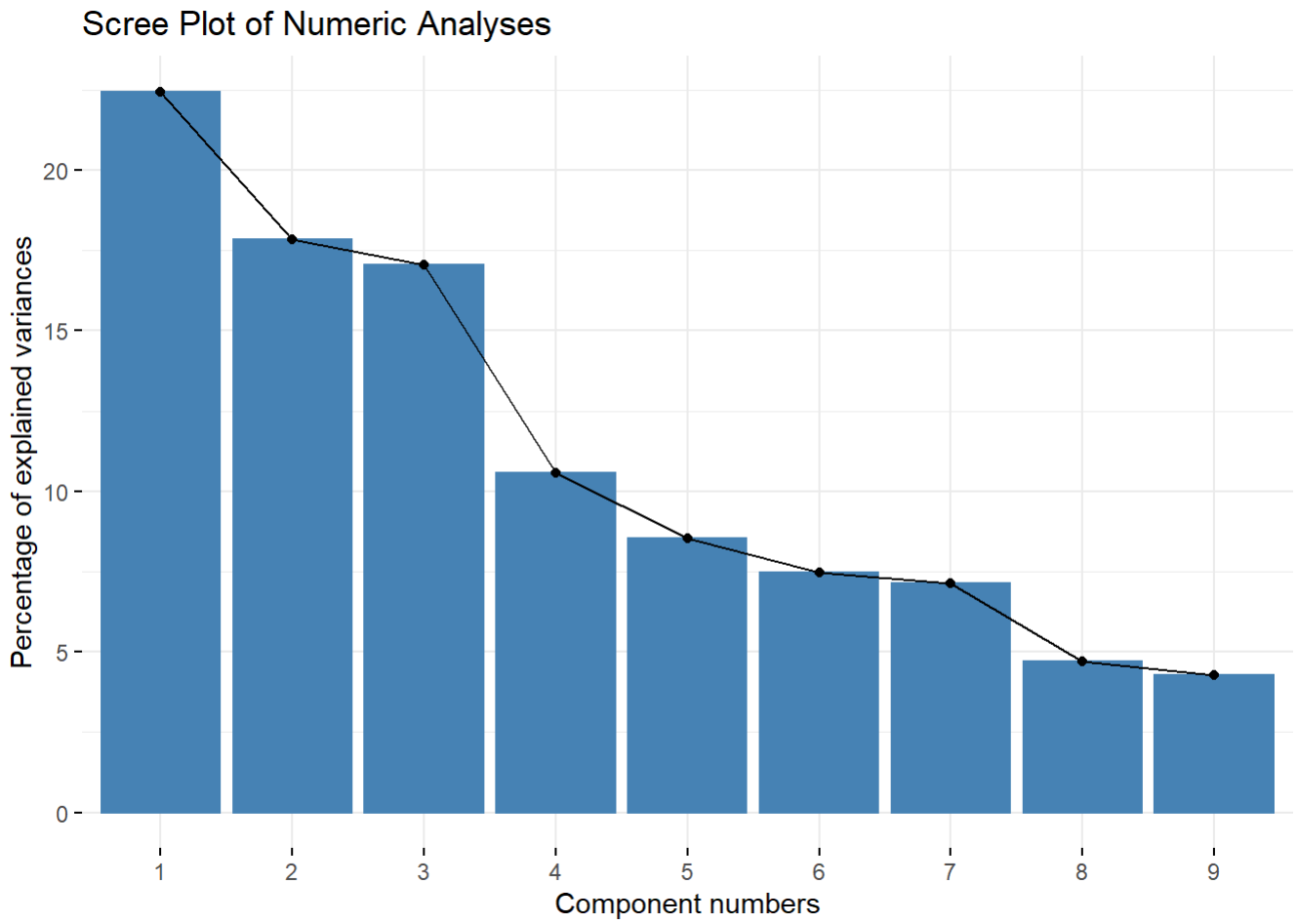
```
## Standard deviations (1, .., p=9):
## [1] 1.4212822 1.2670777 1.2383384 0.9757055 0.8770357 0.8196314 0.8004562
## [8] 0.6503532 0.6199282
##
## Rotation (n x k) = (9 x 9):
##                  PC1          PC2        PC3         PC4         PC5
## illness   -0.5601293  0.037006059 -0.2236361  0.16403673 -0.09547867
## injury    -0.1885349 -0.452999140  0.5071509  0.02917384  0.01630140
## illdays   -0.5473310 -0.041455726 -0.2077741  0.05069248 -0.16579253
## actdays   -0.1889304 -0.455765969  0.5060518  0.01339135 -0.01152437
## commune   -0.1341267  0.458145882  0.3560837 -0.06401366  0.25177099
## lnhhexp.x  0.1724541 -0.407139471 -0.3096811  0.02416622  0.75198933
## pharvis   -0.4741160  0.009870844 -0.1639901  0.12626000  0.40291951
## age       -0.1476331 -0.166722953 -0.1693660 -0.94287067 -0.07483109
## educ       0.1453679 -0.424766949 -0.3394860  0.24466986 -0.40764900
##                  PC6           PC7          PC8          PC9
## illness    0.13902422 -0.1839253619  0.516884952  0.527751842
## injury    -0.01678484 -0.0067554462  0.526265289 -0.472911148
## illdays    0.24140827 -0.3819285156 -0.448274541 -0.469452652
## actdays    0.01197098 -0.0001646119 -0.486645532  0.513044709
## commune   -0.55099697 -0.5233158056 -0.011288310  0.015680987
## lnhhexp.x  0.10337960 -0.3606187827  0.018134951  0.038217275
## pharvis   -0.40190452  0.6196627492 -0.118865454 -0.101488631
## age       -0.14607800  0.0099055478  0.059009413  0.046896292
## educ      -0.65178083 -0.1796413809 -0.005783176  0.005168698
```

```
## 2.) Determining Principle Components
library(factoextra)
```

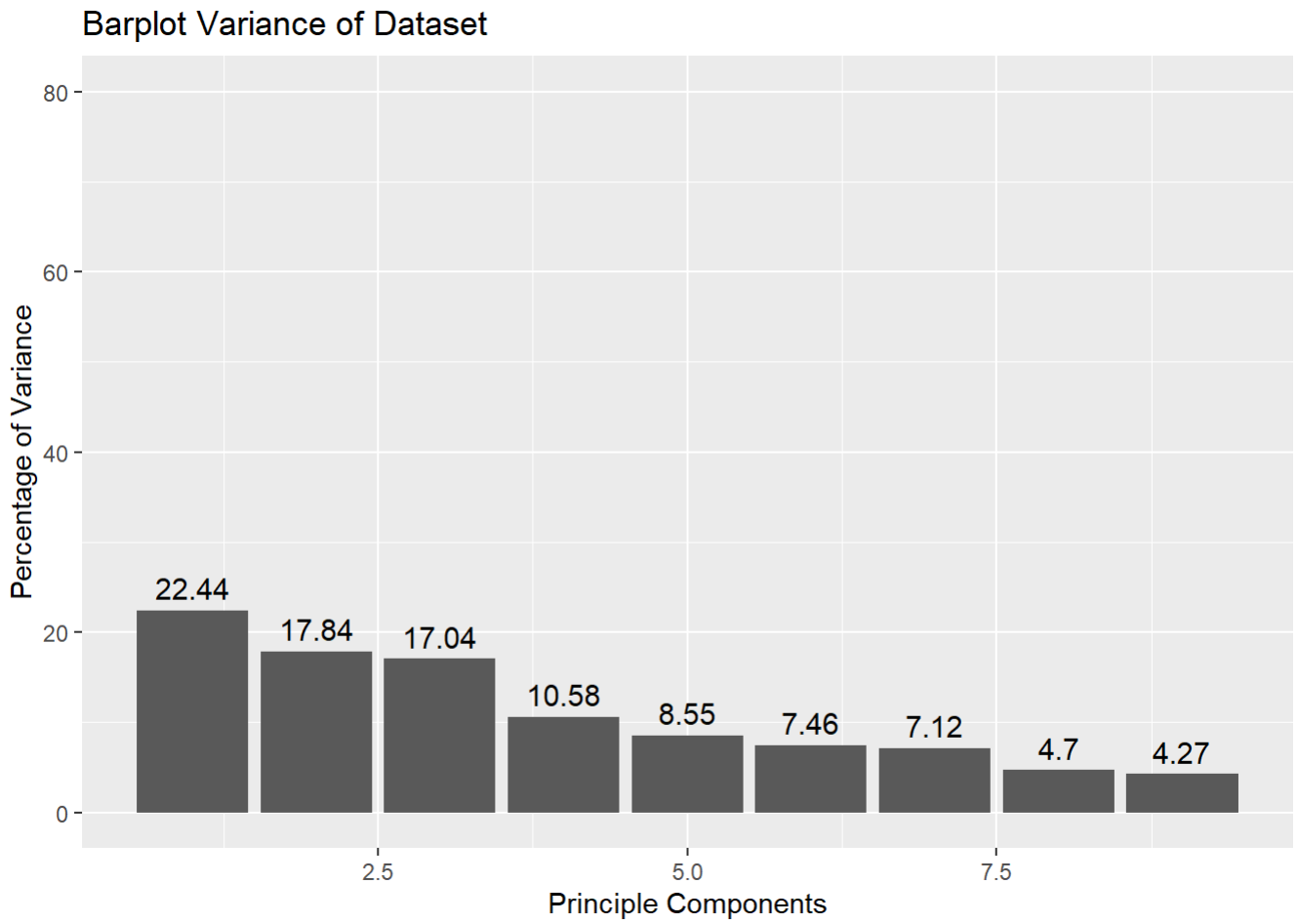
```
## Warning: package 'factoextra' was built under R version 3.6.3
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

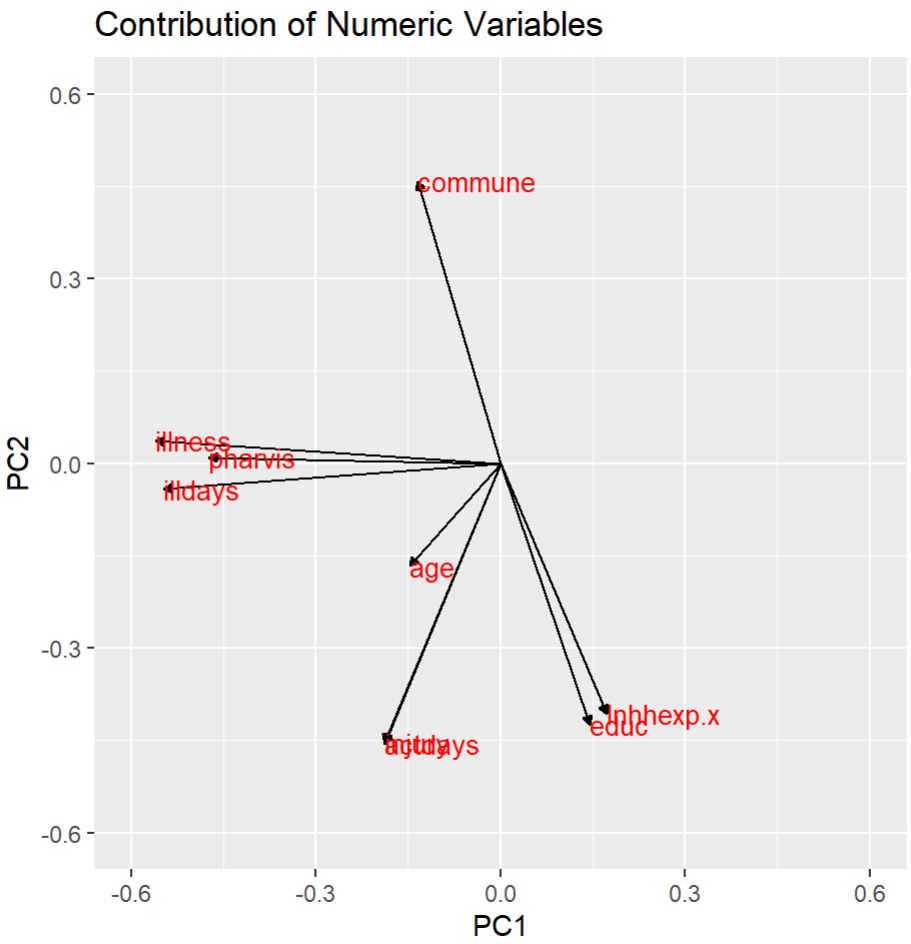
```
fviz_screeplot(VN_PCA) + ggtitle("Scree Plot of Numeric Analyses") + xlab("Component numbers")
```




```
## A barchart containing the variance percentages of each principle component was conducted to f
urther reaffirm this hypothesis. In the first step, standard deviation was factored in order to
 calculate the percentage of variance.
percent <- 100* (VN_PCA$sdev^2 / sum(VN_PCA$sdev^2))
perc_data <- data.frame(percent = percent, PC = 1:length(percent))
## Barchart conveying variance percentages for principle components
ggplot(perc_data, aes(x = PC, y = percent)) +
  geom_col() +
  ggtitle("Barplot Variance of Dataset") +
  ylab("Percentage of Variance") +
  xlab("Principle Components") +
  geom_text(aes(label = round(percent, 2)), size = 4, vjust = -0.5) +
  ylim(0, 80)
```



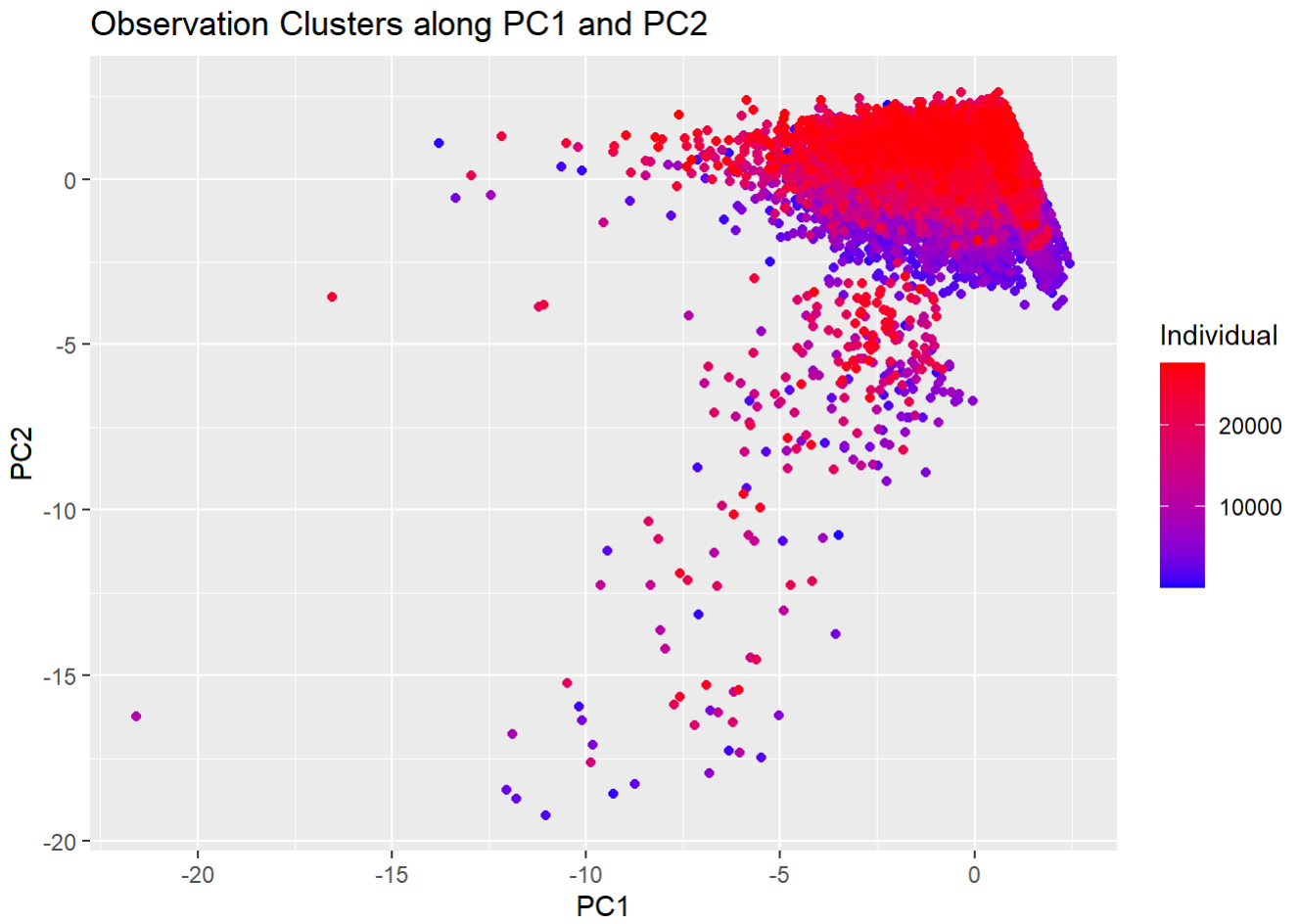

```
## 3.) Principle Component Contribution
## Rotation data frame was established prior to conducting ggplot.
rotation_data <- data.frame(
  VN_PCA$rotation,
  variable = row.names(VN_PCA$rotation))
## Arrow style was reformatted to be used in ggplot.
arrow_style <- arrow(length = unit(0.05, "inches"), type = "closed")
## A ggplot graph was created to observe the contributions of each variables and compare them wi
th eachother.
ggplot(rotation_data) +
  geom_segment(aes(xend = PC1, yend = PC2), x = 0, y = 0, arrow = arrow_style) +
  geom_text(aes(x = PC1, y = PC2, label = variable), hjust = 0, size = 3.5, color = "red") +
  ggtitle("Contribution of Numeric Variables") +
xlim(-0.6, 0.6) +
  ylim(-0.6, 0.6) +
  coord_fixed()
```




```
## 4.) Data Visualization with Cluster
library(ggplot2)
library(cluster)
## The principle components are merged with individual observations from the original dataset.
VN_PCA
```

```
## Standard deviations (1, .., p=9):
## [1] 1.4212822 1.2670777 1.2383384 0.9757055 0.8770357 0.8196314 0.8004562
## [8] 0.6503532 0.6199282
##
## Rotation (n x k) = (9 x 9):
##                  PC1          PC2         PC3         PC4         PC5
## illness   -0.5601293  0.037006059 -0.2236361  0.16403673 -0.09547867
## injury    -0.1885349 -0.452999140  0.5071509  0.02917384  0.01630140
## illdays   -0.5473310 -0.041455726 -0.2077741  0.05069248 -0.16579253
## actdays   -0.1889304 -0.455765969  0.5060518  0.01339135 -0.01152437
## commune   -0.1341267  0.458145882  0.3560837 -0.06401366  0.25177099
## lnhhexp.x  0.1724541 -0.407139471 -0.3096811  0.02416622  0.75198933
## pharvis   -0.4741160  0.009870844 -0.1639901  0.12626000  0.40291951
## age       -0.1476331 -0.166722953 -0.1693660 -0.94287067 -0.07483109
## educ       0.1453679 -0.424766949 -0.3394860  0.24466986 -0.40764900
##                  PC6           PC7          PC8          PC9
## illness    0.13902422 -0.1839253619  0.516884952  0.527751842
## injury    -0.01678484 -0.0067554462  0.526265289 -0.472911148
## illdays    0.24140827 -0.3819285156 -0.448274541 -0.469452652
## actdays    0.01197098 -0.0001646119 -0.486645532  0.513044709
## commune   -0.55099697 -0.5233158056 -0.011288310  0.015680987
## lnhhexp.x  0.10337960 -0.3606187827  0.018134951  0.038217275
## pharvis   -0.40190452  0.6196627492 -0.118865454 -0.101488631
## age       -0.14607800  0.0099055478  0.059009413  0.046896292
## educ      -0.65178083 -0.1796413809 -0.005783176  0.005168698
```

```
PCA_1 <- as.data.frame(VN_PCA[["x"]])
PCA_2 <- bind_cols(VN_combined1['Individual'],PCA_1)
## A scatterplot was constructed to convey the clusters of observations along principle componen
t 1 and 2.
ggplot(PCA_2, aes(x = PC1, y = PC2, color = Individual)) +
  scale_color_gradient(low="blue",high="red") +
  geom_point() +
  ggtitle("Observation Clusters along PC1 and PC2")
```



A PCA test was conducted in order to see whether there were groups of observations that were similar to each other. Likewise, a Principle component would reflect a direction in which variance occurred. This graph focuses on the first two principle components, bearing the most variance within the data. There are three distinguishable clusters within the graph (top right, mid right, and bottom center). These clusters represents the number of individuals that are grouped together (strength indicated by coloring) and supports that there is indeed correlation between numeric variables in the dataset.

A scree plot was constructed in order to determine the necessary principle components by using Kaizer's Rule. Although Kaizer's rule suggested to retain all components with an eigenvalue higher than one, the trendline of the scree plot curved dramatically at principle component number 3, meaning that most of the variance in the dataset is attributed to the first three principle components. Likewise, three is the ideal number of clustering as indicated by both the scree plot and the bar plot.

The arrow plot in step 3 indicates the contribution of variance in all numeric variables. Variables "commune", "illness", and "illdays" displayed the most contribution when compared to the rest. Their high contribution may suggest their strong correlation with PCs 1 and 2, and thus associated with high variance.