R Notebook

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✔ ggplot2 3.3.5     ✔ purrr   0.3.4
## ✔ tibble  3.1.6     ✔ dplyr   1.0.8
## ✔ tidyr   1.2.0     ✔ stringr 1.4.0
## ✔ readr   2.1.2     ✔ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
churn_df <- read.csv("/Users/wyattromero/Downloads/Principal-Component-Analysis/churn_clean.csv")
head(churn_df)
```

	CaseOrder <dbl>	Customer_id <chr>	Interaction <chr>	UID <chr>
1	1	K409198	aa90260b-4141-4a24-8e36-b04ce1f4f77b	e885b299883d4f9fb18e39c75155d990
2	2	S120509	fb76459f-c047-4a9d-8af9-e0f7d4ac2524	f2de8bef964785f41a295929830fb8a
3	3	K191035	344d114c-3736-4be5-98f7-c72c281e2d35	f1784cfa9f6d92ae816197eb175d3c71
4	4	D90850	abfa2b40-2d43-4994-b15a-989b8c79e311	dc8a365077241bb5cd5ccd305136b05e
5	5	K662701	68a861fd-0d20-4e51-a587-8a90407ee574	aabb64a116e83fc4befc1fbab1663f9
6	6	W303516	2b451d12-6c2b-4cea-a295-ba1d6bced078	97598fd95658c80505046c1dd312994

6 rows | 1-5 of 51 columns

```
str(churn_df)
```

```
## 'data.frame':    10000 obs. of  50 variables:
##  $ CaseOrder         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Customer_id       : chr  "K409198" "S120509" "K191035" "D90850" ...
##  $ Interaction       : chr  "aa90260b-4141-4a24-8e36-b04ce1f4f77b" "fb76459f-c047-4a9d-8af9-e0f7d4ac2524" "3
44d114c-3736-4be5-98f7-c72c281e2d35" "abfa2b40-2d43-4994-b15a-989b8c79e311" ...
##  $ UID               : chr  "e885b299883d4f9fb18e39c75155d990" "f2de8bef964785f41a295929830fb8a" "f1784cfa9
f6d92ae816197eb175d3c71" "dc8a365077241bb5cd5ccd305136b05e" ...
##  $ City              : chr  "Point Baker" "West Branch" "Yamhill" "Del Mar" ...
##  $ State             : chr  "AK" "MI" "OR" "CA" ...
##  $ County            : chr  "Prince of Wales-Hyder" "Ogemaw" "Yamhill" "San Diego" ...
##  $ Zip               : int  99927 48661 97148 92014 77461 31030 37847 73109 34771 45237 ...
##  $ Lat               : num  56.3 44.3 45.4 33 29.4 ...
##  $ Lng               : num  -133.4 -84.2 -123.2 -117.2 -95.8 ...
##  $ Population        : int  38 10446 3735 13863 11352 17701 2535 23144 17351 20193 ...
##  $ Area              : chr  "Urban" "Urban" "Urban" "Suburban" ...
##  $ TimeZone          : chr  "America/Sitka" "America/Detroit" "America/Los_Angeles" "America/Los_Angeles"
...
##  $ Job               : chr  "Environmental health practitioner" "Programmer, multimedia" "Chief Financial Of
ficer" "Solicitor" ...
##  $ Children          : int  0 1 4 1 0 3 0 2 2 1 ...
##  $ Age               : int  68 27 50 48 83 83 79 30 49 86 ...
##  $ Income            : num  28562 21705 9610 18925 40074 ...
##  $ Marital           : chr  "Widowed" "Married" "Widowed" "Married" ...
##  $ Gender            : chr  "Male" "Female" "Female" "Male" ...
##  $ Churn             : chr  "No" "Yes" "No" "No" ...
##  $ Outage_sec_perweek  : num  7.98 11.7 10.75 14.91 8.15 ...
##  $ Email             : int  10 12 9 15 16 15 10 16 20 18 ...
##  $ Contacts          : int  0 0 0 2 2 3 0 0 2 1 ...
##  $ Yearly_equip_failure: int  1 1 1 0 1 1 1 0 3 0 ...
##  $ Techie            : chr  "No" "Yes" "Yes" "Yes" ...
##  $ Contract          : chr  "One year" "Month-to-month" "Two Year" "Two Year" ...
##  $ Port_modem        : chr  "Yes" "No" "Yes" "No" ...
##  $ Tablet            : chr  "Yes" "Yes" "No" "No" ...
##  $ InternetService   : chr  "Fiber Optic" "Fiber Optic" "DSL" "DSL" ...
##  $ Phone             : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Multiple          : chr  "No" "Yes" "Yes" "No" ...
##  $ OnlineSecurity    : chr  "Yes" "Yes" "No" "Yes" ...
##  $ OnlineBackup      : chr  "Yes" "No" "No" "No" ...
##  $ DeviceProtection  : chr  "No" "No" "No" "No" ...
##  $ TechSupport       : chr  "No" "Yes" "No" "No" ...
##  $ StreamingTV       : chr  "No" "Yes" "No" "Yes" ...
##  $ StreamingMovies   : chr  "Yes" "Yes" "Yes" "No" ...
##  $ PaperlessBilling  : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ PaymentMethod     : chr  "Credit Card (automatic)" "Bank Transfer(automatic)" "Credit Card (automatic)"
"Mailed Check" ...
##  $ Tenure            : num  6.8 1.16 15.75 17.09 1.67 ...
##  $ MonthlyCharge     : num  172 243 160 120 150 ...
##  $ Bandwidth_GB_Year : num  905 801 2055 2165 271 ...
##  $ Item1             : int  5 3 4 4 4 3 6 2 5 2 ...
##  $ Item2             : int  5 4 4 4 4 3 5 2 4 2 ...
##  $ Item3             : int  5 3 2 4 4 3 6 2 4 2 ...
##  $ Item4             : int  3 3 4 2 3 2 4 5 3 2 ...
##  $ Item5             : int  4 4 4 5 4 4 1 2 4 5 ...
##  $ Item6             : int  4 3 3 4 4 3 5 3 3 2 ...
##  $ Item7             : int  3 4 3 3 4 3 5 4 4 3 ...
##  $ Item8             : int  4 4 3 3 5 3 5 4 3 ...
```

```
churn_df <- select(churn_df, c('Age', 'Income', 'Outage_sec_perweek', 'MonthlyCharge', 'Bandwidth_GB_Year', 'Tenu
re', 'Churn'))
head(churn_df)
```

	A… <int>	Income <dbl>	Outage_sec_perweek <dbl>	MonthlyCharge <dbl>	Bandwidth_GB_Year <dbl>	Tenure <dbl>	Churn <chr>
1	68	28561.99	7.978323	172.4555	904.5361	6.795513	No
2	27	21704.77	11.699080	242.6326	800.9828	1.156681	Yes
3	50	9609.57	10.752800	159.9476	2054.7070	15.754144	No
4	48	18925.23	14.913540	119.9568	2164.5794	17.087227	No
5	83	40074.19	8.147417	149.9483	271.4934	1.670972	Yes
6	83	22660.20	8.420993	185.0077	1039.3580	7.000994	No

6 rows

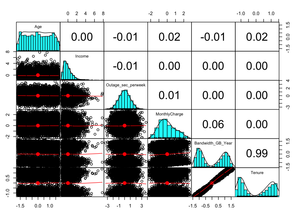
```
churn_df_scaled <- scale(churn_df[1:6])
head(churn_df_scaled)
```

```
##             Age       Income Outage_sec_perweek MonthlyCharge Bandwidth_GB_Year
## [1,]  0.7208892 -0.398757801       -0.6799436  -0.003942362      -1.1384301
## [2,] -1.2598942 -0.641922349        0.5703026   1.638244379      -1.1858165
## [3,] -0.1487230 -1.078831417        0.2523344  -0.295210056      -0.6121071
## [4,] -0.2453466 -0.749487888        1.6504233  -1.226459744      -0.5618291
## [5,]  1.4455660  0.009477447       -0.6231249  -0.528959300      -1.4281131
## [6,]  1.4455660 -0.608041752       -0.5311979   0.288355463      -1.0767351
##            Tenure
## [1,] -1.0486938
## [2,] -1.2619381
## [3,] -0.7099043
## [4,] -0.6594910
## [5,] -1.2424891
## [6,] -1.0409231
```

```
write.csv(churn_df_scaled, "Churn_prepared_D212_2.csv", row.names =TRUE)
```

```
set.seed(111)
ind <- sample(2, nrow(churn_df_scaled),
              replace = TRUE,
              prob = c(0.8, 0.2))
training <- churn_df_scaled[ind==1,]
testing <- churn_df_scaled[ind==2,]
```

```
pairs.panels(training[,-7],
             gap = 0,
             bg = c("red", "yellow", "blue") [churn_df$Churn],
             pch=21)
```




```
pc <- prcomp(training[,-7],
             center = TRUE,
             scale. = TRUE)
attributes(pc)
```

```
## $names
## [1] "sdev"     "rotation" "center"   "scale"    "x"
##
## $class
## [1] "prcomp"
```

```
pc$center
```

```
##                 Age              Income Outage_sec_perweek       MonthlyCharge
##       -0.0070502429       -0.0009029114        0.0015437022       -0.0011858877
##  Bandwidth_GB_Year              Tenure
##       -0.0101698053       -0.0102591866
```

```
pc$scale
```

```
##                 Age              Income Outage_sec_perweek       MonthlyCharge
##           0.9946660           1.0011141           0.9984468           0.9966557
##  Bandwidth_GB_Year              Tenure
##           0.9986297           1.0003683
```

```
print(pc)
```

```
## Standard deviations (1, .., p=6):
## [1] 1.411745623 1.010943393 1.007334732 0.996700656 0.985334467 0.076831111
##
## Rotation (n x k) = (6 x 6):
##                              PC1         PC2         PC3         PC4
## Age                  0.001201900 -0.350804215  0.685031150 0.347862401
## Income               0.002328188 -0.532644714 -0.080265021 -0.797382797
## Outage_sec_perweek  -0.002910350  0.730852291  0.030035144 -0.278938869
## MonthlyCharge        0.037592291  0.242358393  0.722724429 -0.605439435
## Bandwidth_GB_Year    0.707187004  0.008770112 -0.007646915 -0.007873523
## Tenure               0.706015599 -0.016322801 -0.031600938  0.030361933
##                              PC5         PC6
## Age                  0.523492815  0.022603244
## Income               0.272706737 -0.001317134
## Outage_sec_perweek   0.622202961  0.000561058
## MonthlyCharge       -0.501114052 -0.044969209
## Bandwidth_GB_Year   -0.010793171  0.706804579
## Tenure               0.033824881 -0.705614698
```

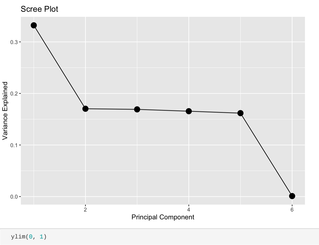
```
summary(pc)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5    PC6
## Standard deviation     1.4117 1.0109 1.0073 0.9967 0.9853 0.07683
## Proportion of Variance 0.3322 0.1703 0.1691 0.1656 0.1618 0.00098
## Cumulative Proportion  0.3322 0.5025 0.6716 0.8372 0.9990 1.00000
```

```
variance = pc$sdev^2 / sum(pc$sdev^2)
variance
```

```
## [1] 0.3321707668 0.1703346056 0.1691247701 0.1655686677 0.1618173533
## [6] 0.0009838365
```

```
qplot(c(1:6), variance) +
  geom_line() +
  geom_point(size=4) +
  xlab("Principal Component") +
  ylab("Variance Explained") +
  ggtitle("Scree Plot")
```




```
ylim(0, 1)
```

```
## <ScaleContinuousPosition>
##  Range:
##  Limits:    0 --    1
```