

DS

```
# getting clinical data for TCGA-BRCA cohort -----
clinical_brca <- GDCquery_clinic("TCGA-ESCA")
any(colnames(clinical_brca) %in% c("vital_status", "days_to_last_follow_up", "days_to_death"))

## [1] TRUE

which(colnames(clinical_brca) %in% c("vital_status", "days_to_last_follow_up", "days_to_death"))

## [1]  9 43 48

clinical_brca[,c(9,43,48)]

##      days_to_last_follow_up vital_status days_to_death
## 1             1378         Dead         1405
## 2             2069         Alive          NA
## 3              NA         Dead          784
## 4              NA         Dead           9
## 5              NA         Dead         272
## 6              NA         Dead         157
## 7              NA         Dead          96
## 8             107         Dead         730
## 9             383         Alive          NA
## 10            1071         Alive          NA
## 11              NA         Dead         855
## 12             112         Dead         801
## 13             660         Alive          NA
## 14              2         Dead         180
## 15             NA         Dead         217
## 16             64         Alive          NA
## 17            1007         Alive          NA
## 18             NA         Dead         480
## 19             NA         Dead         650
## 20            124         Dead        1458
## 21             40         Alive          NA
## 22             NA         Dead           0
## 23             NA         Dead         556
## 24            471         Alive          NA
## 25             81         Alive          NA
## 26             NA         Dead         112
## 27            1590         Alive          NA
## 28             272         Alive          NA
## 29            320         Alive          NA
## 30            408         Alive          NA
## 31            375         Alive          NA
```

## 32	NA	Dead	1361
## 33	391	Alive	NA
## 34	500	Alive	NA
## 35	388	Alive	NA
## 36	383	Alive	NA
## 37	408	Alive	NA
## 38	1168	Alive	NA
## 39	30	Dead	283
## 40	992	Alive	NA
## 41	452	Alive	NA
## 42	4	Dead	303
## 43	NA	Dead	279
## 44	609	Alive	NA
## 45	NA	Dead	694
## 46	NA	Dead	410
## 47	0	Dead	26
## 48	407	Alive	NA
## 49	225	Alive	NA
## 50	375	Alive	NA
## 51	3	Dead	681
## 52	767	Alive	NA
## 53	412	Alive	NA
## 54	NA	Dead	1263
## 55	101	Alive	NA
## 56	NA	Dead	378
## 57	NA	Dead	42
## 58	407	Alive	NA
## 59	1254	Alive	NA
## 60	NA	Dead	393
## 61	11	Dead	390
## 62	441	Alive	NA
## 63	NA	Dead	951
## 64	NA	Dead	386
## 65	NA	Dead	232
## 66	234	Alive	NA
## 67	5	Dead	136
## 68	NA	Dead	424
## 69	1	Dead	567
## 70	620	Alive	NA
## 71	825	Alive	NA
## 72	1688	Alive	NA
## 73	694	Alive	NA
## 74	920	Alive	NA
## 75	612	Alive	NA
## 76	273	Alive	NA
## 77	NA	Dead	610
## 78	785	Alive	NA
## 79	245	Dead	330
## 80	NA	Dead	247
## 81	1641	Alive	NA
## 82	366	Alive	NA
## 83	370	Alive	NA
## 84	632	Alive	NA
## 85	1012	Alive	NA

## 86	769	Dead	960
## 87	NA	Dead	214
## 88	NA	Dead	226
## 89	1094	Alive	NA
## 90	NA	Dead	193
## 91	401	Alive	NA
## 92	379	Alive	NA
## 93	467	Alive	NA
## 94	372	Alive	NA
## 95	437	Alive	NA
## 96	NA	Dead	154
## 97	3714	Alive	NA
## 98	167	Alive	NA
## 99	NA	Dead	213
## 100	342	Alive	NA
## 101	467	Alive	NA
## 102	NA	Dead	149
## 103	365	Alive	NA
## 104	218	Alive	NA
## 105	1060	Alive	NA
## 106	366	Alive	NA
## 107	8	Dead	118
## 108	70	Alive	NA
## 109	402	Alive	NA
## 110	477	Alive	NA
## 111	501	Alive	NA
## 112	16	Alive	NA
## 113	80	Alive	NA
## 114	14	Dead	142
## 115	NA	Dead	484
## 116	NA	Dead	24
## 117	731	Alive	NA
## 118	380	Alive	NA
## 119	96	Alive	NA
## 120	NA	Dead	236
## 121	472	Alive	NA
## 122	NA	Dead	104
## 123	NA	Dead	180
## 124	NA	Dead	243
## 125	318	Dead	318
## 126	282	Alive	NA
## 127	378	Alive	NA
## 128	882	Alive	NA
## 129	384	Alive	NA
## 130	NA	Dead	283
## 131	385	Alive	NA
## 132	NA	Dead	557
## 133	375	Alive	NA
## 134	375	Alive	NA
## 135	104	Alive	NA
## 136	705	Alive	NA
## 137	1441	Alive	NA
## 138	NA	Dead	764
## 139	1837	Alive	NA

## 140	1025	Alive	NA
## 141	79	Alive	NA
## 142	NA	Dead	1599
## 143	238	Alive	NA
## 144	2	Dead	351
## 145	NA	Dead	494
## 146	NA	Dead	558
## 147	4	Alive	NA
## 148	554	Alive	NA
## 149	608	Alive	NA
## 150	639	Alive	NA
## 151	402	Alive	NA
## 152	NA	Dead	88
## 153	NA	Dead	161
## 154	NA	Dead	435
## 155	NA	Dead	2134
## 156	373	Alive	NA
## 157	NA	Dead	47
## 158	447	Dead	496
## 159	NA	Dead	1402
## 160	1271	Alive	NA
## 161	768	Alive	NA
## 162	824	Alive	NA
## 163	265	Alive	NA
## 164	403	Alive	NA
## 165	315	Alive	NA
## 166	NA	Dead	231
## 167	712	Alive	NA
## 168	549	Alive	NA
## 169	NA	Dead	81
## 170	383	Alive	NA
## 171	NA	Dead	987
## 172	400	Alive	NA
## 173	143	Alive	NA
## 174	NA	Dead	1781
## 175	191	Alive	NA
## 176	401	Alive	NA
## 177	NA	Dead	128
## 178	114	Alive	NA
## 179	NA	Dead	763
## 180	518	Alive	NA
## 181	92	Alive	NA
## 182	11	Alive	NA
## 183	NA	Dead	553
## 184	551	Dead	600
## 185	391	Alive	NA

```
# looking at some variables associated with survival
table(clinical_brca$vital_status)
```

```
##
## Alive Dead
## 108 77
```

```

# change certain values the way they are encoded
clinical_brca$deceased <- ifelse(clinical_brca$vital_status == "Alive", FALSE, TRUE)

# create an "overall survival" variable that is equal to days_to_death
# for dead patients, and to days_to_last_follow_up for patients who
# are still alive
clinical_brca$overall_survival <- ifelse(clinical_brca$vital_status == "Alive",
                                         clinical_brca$days_to_last_follow_up,
                                         clinical_brca$days_to_death)

```

```

# build a query to get gene expression data for entire cohort
query_brca_all = GDCquery(
  project = "TCGA-ESCA",
  data.category = "Transcriptome Profiling", # parameter enforced by GDCquery
  experimental.strategy = "RNA-Seq",
  workflow.type = "STAR - Counts",
  data.type = "Gene Expression Quantification",
  access = "open")

```

```

## -----

## o GDCquery: Searching in GDC database

## -----

## Genome of reference: hg38

## -----

## oo Accessing GDC. This might take a while...

## -----

## ooo Project: TCGA-ESCA

## -----

## oo Filtering results

## -----

## ooo By access

## ooo By experimental.strategy

## ooo By data.type

## ooo By workflow.type

```

```

## -----

## oo Checking data

## -----

## ooo Checking if there are duplicated cases

## ooo Checking if there are results for the query

## -----

## o Preparing output

## -----

output_brca <- getResults(query_brca_all)
# get primary tissue sample barcodes
tumor <- output_brca$cases#[1:50]

# # get gene expression data from primary tumors
query_brca <- GDCquery(
  project = "TCGA-ESCA",
  data.category = "Transcriptome Profiling", # parameter enforced by GDCquery
  experimental.strategy = "RNA-Seq",
  workflow.type = "STAR - Counts",
  data.type = "Gene Expression Quantification",
  sample.type = c("Primary Tumor", "Solid Tissue Normal"),
  access = "open",
  barcode = tumor)

## -----

## o GDCquery: Searching in GDC database

## -----

## Genome of reference: hg38

## -----

## oo Accessing GDC. This might take a while...

## -----

## ooo Project: TCGA-ESCA

## -----

```

```

## oo Filtering results

## -----

## ooo By access

## ooo By experimental.strategy

## ooo By data.type

## ooo By workflow.type

## ooo By barcode

## ooo By sample.type

## -----

## oo Checking data

## -----

## ooo Checking if there are duplicated cases

## ooo Checking if there are results for the query

## -----

## o Preparing output

## -----

# download data
GDCdownload(query_brca)

## Downloading data for project TCGA-ESCA

## Of the 197 files for download 197 already exist.

## All samples have been already downloaded

library(SummarizedExperiment)

## Loading required package: MatrixGenerics

## Loading required package: matrixStats

```

```

##
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars

## Loading required package: GenomicRanges

## Loading required package: stats4

## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

```



```

## Loading required package: IRanges

## Loading required package: GenomeInfoDb

## Loading required package: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase)", and for packages 'citation("pkgname)".

##
## Attaching package: 'Biobase'

## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians

## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians

# get counts
tcga_brca_data <- GDCprepare(query_brca, summarizedExperiment = T)

## |                                | 0% |

## Starting to add information to samples

## => Add clinical information to samples

## => Adding TCGA molecular information from marker papers

## => Information will have prefix 'paper_'

## esca subtype information from:doi:10.1038/nature20805

## Available assays in SummarizedExperiment :
##     => unstranded
##     => stranded_first
##     => stranded_second
##     => tpm_unstrand
##     => fpkm_unstrand
##     => fpkm_uq_unstrand

brca_matrix <- assay(tcga_brca_data)
brca_matrix[1:10,1:10]

```

##	TCGA-L5-A40M-01A-11R-A260-31	TCGA-L5-A43J-01A-12R-A24K-31
## ENSG000000000003.15	2159	4526
## ENSG000000000005.6	2	0
## ENSG000000000419.13	2742	5195
## ENSG000000000457.14	1909	837
## ENSG000000000460.17	1647	608
## ENSG000000000938.13	323	406
## ENSG000000000971.16	4274	5214
## ENSG00000001036.14	2151	4815
## ENSG00000001084.13	9737	3023
## ENSG00000001167.14	4649	1552
##	TCGA-LN-A49K-01A-11R-A24K-31	TCGA-L5-A40H-01A-11R-A260-31
## ENSG000000000003.15	3786	6229
## ENSG000000000005.6	4	13
## ENSG000000000419.13	3441	12243
## ENSG000000000457.14	1821	1851
## ENSG000000000460.17	1265	1619
## ENSG000000000938.13	912	265
## ENSG000000000971.16	8294	1696
## ENSG00000001036.14	2680	4474
## ENSG00000001084.13	36988	11008
## ENSG00000001167.14	3393	4711
##	TCGA-JY-A6FE-01A-11R-A336-31	TCGA-VR-A8Q7-01A-11R-A37I-31
## ENSG000000000003.15	3291	2744
## ENSG000000000005.6	2	3
## ENSG000000000419.13	6489	3434
## ENSG000000000457.14	483	797
## ENSG000000000460.17	810	1676
## ENSG000000000938.13	628	365
## ENSG000000000971.16	4751	1094
## ENSG00000001036.14	4416	4432
## ENSG00000001084.13	9801	2420
## ENSG00000001167.14	1753	2310
##	TCGA-IG-A7DP-01A-31R-A336-31	TCGA-L5-A43M-01A-11R-A24K-31
## ENSG000000000003.15	852	1547
## ENSG000000000005.6	4	2
## ENSG000000000419.13	1808	2886
## ENSG000000000457.14	743	1358
## ENSG000000000460.17	246	529
## ENSG000000000938.13	1819	430
## ENSG000000000971.16	9678	3608
## ENSG00000001036.14	2363	4198
## ENSG00000001084.13	1931	4371
## ENSG00000001167.14	1036	2503
##	TCGA-2H-A9GF-01A-11R-A37I-31	TCGA-R6-A6XG-01B-11R-A336-31
## ENSG000000000003.15	2967	2321
## ENSG000000000005.6	0	3
## ENSG000000000419.13	2563	2969
## ENSG000000000457.14	1229	1352
## ENSG000000000460.17	837	869
## ENSG000000000938.13	474	61
## ENSG000000000971.16	1800	527
## ENSG00000001036.14	3822	4683
## ENSG00000001084.13	5207	4129

```
# extract gene and sample metadata from summarizedExperiment object
gene_metadata <- as.data.frame(rowData(tcga_brca_data))
coldata <- as.data.frame(colData(tcga_brca_data))
```

```
# vst transform counts to be used in survival analysis -----
library(DESeq2)
# Setting up countData object
dds <- DESeqDataSetFromMatrix(countData = brca_matrix,
                              colData = coldata,
                              design = ~ 1)

# Removing genes with sum total of 10 reads across all samples
keep <- rowSums(counts(dds)) >= 10
dds <- dds[keep,]
```

```
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:S4Vectors':
##
##     expand
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:Biobase':
##
##     combine
```

```
## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union
```

```
## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect
```

```
## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union
```

```
## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union
```

```
## The following objects are masked from 'package:BiocGenerics':
##
##   combine, intersect, setdiff, union

## The following object is masked from 'package:matrixStats':
##
##   count

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
library(tibble)
# vst
vsd <- vst(dds, blind=FALSE)
brca_matrix_vst <- assay(vsd)
brca_matrix_vst[1:10,1:10]
```

```
##          TCGA-L5-A40M-01A-11R-A260-31 TCGA-L5-A43J-01A-12R-A24K-31
## ENSG000000000003.15          10.758486          12.083141
## ENSG000000000005.6           2.890773           1.990843
## ENSG000000000419.13          11.101946          12.281688
## ENSG000000000457.14          10.581805           9.659773
## ENSG000000000460.17          10.370015           9.203917
## ENSG000000000938.13           8.054563           8.630888
## ENSG000000000971.16          11.740436          12.286947
## ENSG00000001036.14          10.753155          12.172282
## ENSG00000001084.13          12.926446          11.502198
## ENSG00000001167.14          11.861499          10.544080
##          TCGA-LN-A49K-01A-11R-A24K-31 TCGA-L5-A40H-01A-11R-A260-31
## ENSG000000000003.15          11.283802          12.093817
## ENSG000000000005.6           3.133356           4.006536
## ENSG000000000419.13          11.146416          13.067413
## ENSG000000000457.14          10.232801          10.349301
## ENSG000000000460.17           9.711374          10.157355
## ENSG000000000938.13           9.244598           7.596144
## ENSG000000000971.16          12.412696          10.223933
## ENSG00000001036.14          10.787254          11.617409
## ENSG00000001084.13          14.567981          12.914158
## ENSG00000001167.14          11.126221          11.691694
##          TCGA-JY-A6FE-01A-11R-A336-31 TCGA-VR-A8Q7-01A-11R-A37I-31
## ENSG000000000003.15          11.825784          11.361270
## ENSG000000000005.6           3.041129           3.182106
## ENSG000000000419.13          12.803698          11.684005
## ENSG000000000457.14           9.075528           9.588216
## ENSG000000000460.17           9.812870          10.652781
## ENSG000000000938.13           9.449385           8.478920
## ENSG000000000971.16          12.354519          10.041145
## ENSG00000001036.14          12.249194          12.051288
```

## ENSG00000001084.13	13.398089	11.180579
## ENSG00000001167.14	10.919849	11.113700
##	TCGA-IG-A7DP-01A-31R-A336-31	TCGA-L5-A43M-01A-11R-A24K-31
## ENSG00000000003.15	10.285076	10.336477
## ENSG00000000005.6	3.644639	2.908037
## ENSG000000000419.13	11.365688	11.231964
## ENSG000000000457.14	10.088928	10.149717
## ENSG000000000460.17	8.515192	8.805212
## ENSG000000000938.13	11.374413	8.512114
## ENSG000000000971.16	13.782467	11.553139
## ENSG00000001036.14	11.750893	11.771106
## ENSG00000001084.13	11.460365	11.829237
## ENSG00000001167.14	10.565546	11.027279
##	TCGA-2H-A9GF-01A-11R-A37I-31	TCGA-R6-A6XG-01B-11R-A336-31
## ENSG00000000003.15	11.609599	11.372338
## ENSG00000000005.6	1.990843	3.284580
## ENSG000000000419.13	11.399005	11.726625
## ENSG000000000457.14	10.343248	10.595777
## ENSG000000000460.17	9.793176	9.962210
## ENSG000000000938.13	8.982641	6.270712
## ENSG000000000971.16	10.890964	9.248060
## ENSG00000001036.14	11.974101	12.382843
## ENSG00000001084.13	12.419468	12.201490
## ENSG00000001167.14	10.917849	11.972512

```

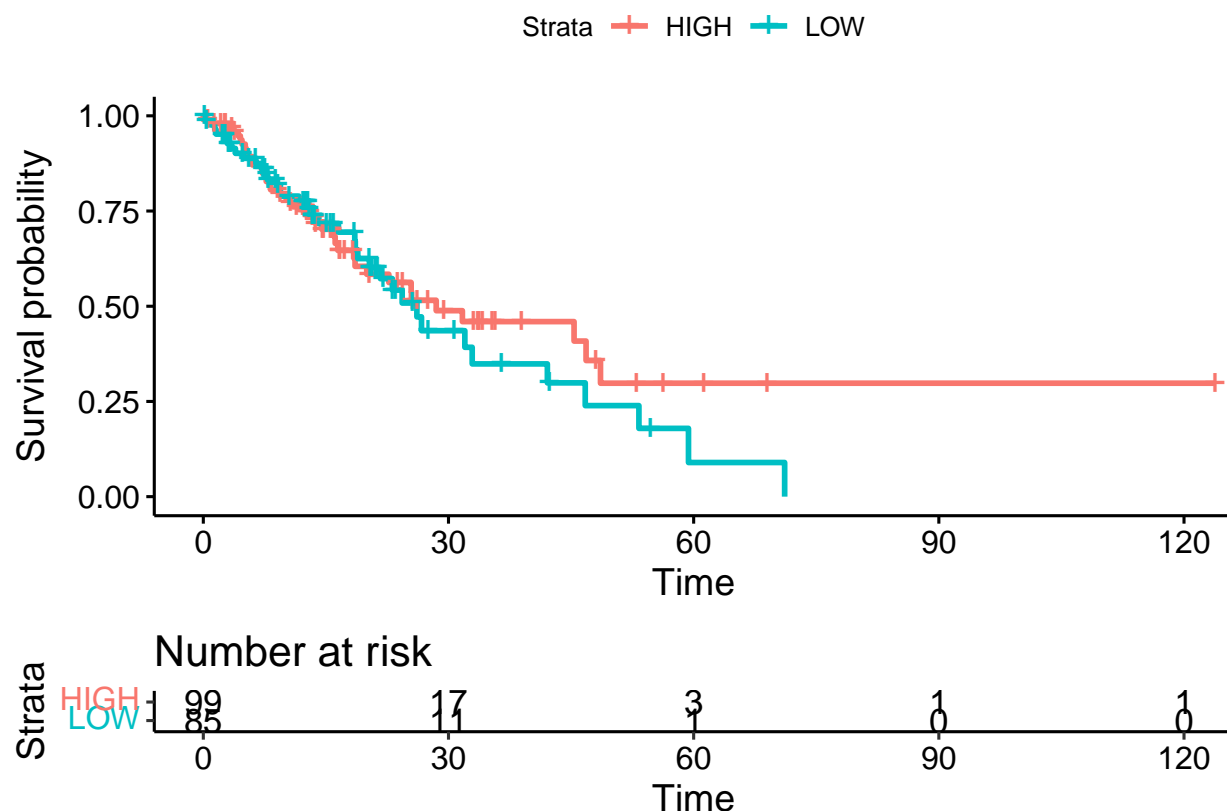
# Get data for TP53 gene and add gene metadata information to it -----
brca_tp53 <- brca_matrix_vst %>%
  as.data.frame() %>%
  rownames_to_column(var = 'gene_id') %>%
  gather(key = 'case_id', value = 'counts', -gene_id) %>%
  left_join(., gene_metadata, by = "gene_id") %>%
  filter(gene_name == "TRIP13")
# get median value
median_value <- median(brca_tp53$counts)

# denote which cases have higher lower expression than median count
brca_tp53$strata <- ifelse(brca_tp53$counts >= median_value, "HIGH", "LOW")

# Add clinical information to brca_tp53
brca_tp53$case_id <- gsub('-01.*', '', brca_tp53$case_id)
brca_tp53 <- merge(brca_tp53, clinical_brca, by.x = 'case_id', by.y = 'submitter_id')
# Convert days to months for overall_survival variable
brca_tp53$overall_survival <- brca_tp53$overall_survival / 30

# fitting survival curve -----
fit <- survfit(Surv(overall_survival, deceased) ~ strata, data = brca_tp53)
ggsurvplot(fit,
  data = brca_tp53,
  risk.table = T)

```



R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

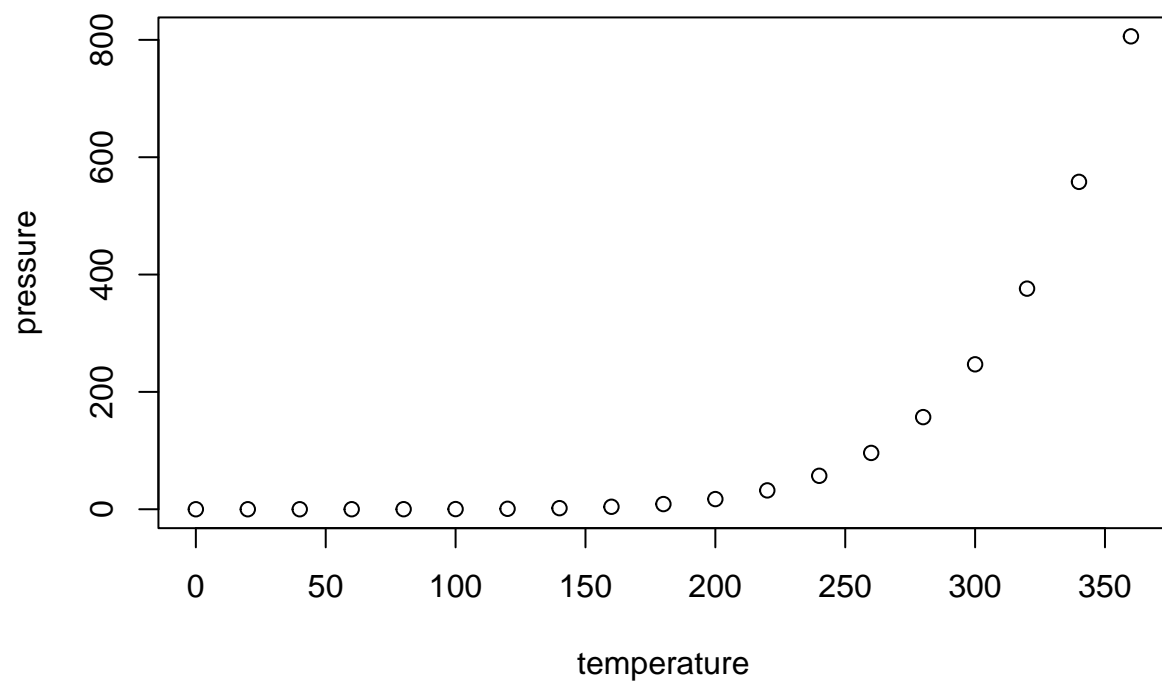
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed      dist
##  Min.   : 4.0    Min.   : 2.00
##  1st Qu.:12.0    1st Qu.: 26.00
##  Median :15.0    Median : 36.00
##  Mean   :15.4    Mean   : 42.98
##  3rd Qu.:19.0    3rd Qu.: 56.00
##  Max.   :25.0    Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.