install.packages("gtsummary")

install.packages("survival")

install.packages("survminer")

install.packages("forestplot")

install.packages("tableone")

install.packages("broom")

install.packages("caret")

install.packages("gridExtra")

install.packages("ggplot2")

install.packages("stringr")

install.packages("gdtools")

install.packages("flextable")

library(readr)

library(dplyr)

library(tidyr)

library(magrittr)

library(gtsummary)

library(survival)

library(purrr)

library(caret)

library(survminer)

library(tableone)

library(broom)

library(gridExtra)

library(ggplot2)

library(stringr)

library(flextable)

library(gdtools)

data <- read\_csv("Desktop/sseer/原始数据/HCC.csv")

new\_data <- data %>%

mutate(Age = `Age recode with single ages and 100+`,

Gender = `Sex`,

Race = `Race recode (W, B, AI, API)`,

Marital\_status = `Marital status at diagnosis`,

Histological\_type = `Histologic Type ICD-O-3`,

Grade = `Grade (thru 2017)`,

Stage = `Derived AJCC Stage Group, 7th ed (2010-2015)`,

T = `Derived AJCC T, 7th ed (2010-2015)`,

N = `Derived AJCC N, 7th ed (2010-2015)`,

M = `Derived AJCC M, 7th ed (2010-2015)`,

AFP = `AFP Pretreatment Interpretation Recode (2010+)`,

Fibrosis\_score = `Fibrosis Score Recode (2010+)`,

Number\_of\_tumors = `Total number of in situ/malignant tumors for patient`,

Tumor\_size = `CS tumor size (2004-2015)`,

Surgery = `RX Summ--Surg Prim Site (1998+)`,

Radiotherapy = `RX Summ--Surg/Rad Seq`,

Chemotherapy = `Chemotherapy recode (yes, no/unk)`,

Survival\_months = `Survival months`,

Status = `Vital status recode (study cutoff used)`) %>%

select( Age, Gender, Race, Marital\_status, Histological\_type, Grade, Stage, T, N, M, AFP,

Fibrosis\_score, Number\_of\_tumors, Tumor\_size, Surgery, Radiotherapy, Chemotherapy, Survival\_months, Status)

filtered\_data <- new\_data %>%

filter(Age >= 18,

Survival\_months != "0000",

Race != "Unknown",

Marital\_status != "Unknown",

Grade != "Unknown",

!(Stage %in% c("IIINOS", "UNK Stage", "NA")),

!(T %in% c("T3NOS", "NA", "TX", "T0")),

!N %in% c("NX"),

!AFP %in% c("Blank(s)", "Borderline; undetermined if positive or negative", "Not documented; Not assessed or unknown if assessed", "Test ordered, results not in chart"),

!Fibrosis\_score %in% c("Unknown; MR statement w/o hist conf; Uncategorized", "Not applicable: Information not collected for this case", "Blank(s)"),

!Tumor\_size %in% c("989", "999"),

Surgery != "90")

filtered\_data <- filtered\_data %>%

  mutate(Age = as.numeric(gsub(" y.\*", "", Age)))

#Age

filtered\_data <- filtered\_data %>%

  mutate(Age = as.numeric(Age)) %>%

  mutate(Age = cut(Age, breaks=c(-Inf,66,77,Inf),labels=c('≤ 66','> 66, ≤ 77','> 77')))

filtered\_data <- filtered\_data %>%

  mutate(Marital\_status = case\_when(

    Marital\_status %in% c("Married (including common law)") ~ "Married",

    Marital\_status %in% c("Single (never married)", "Divorced", "Separated", "Widowed", "Unmarried or Domestic Partner") ~ "Other",

    TRUE ~ Marital\_status))

# Stage

filtered\_data <- filtered\_data %>%

  mutate(Stage = recode(Stage,

                        "I" = "I",

                        "II" = "II",

                        "IIIA" = "III",

                        "IIIB" = "III",

                        "IIIC" = "III",

                        "IVA" = "IV",

                        "IVB" = "IV"))

# Tumor size

filtered\_data <- filtered\_data %>%

  mutate(Tumor\_size = as.numeric(Tumor\_size)) %>%

  mutate(Tumor\_size = cut(Tumor\_size, breaks=c(-Inf, 62, Inf), labels=c('≤ 62 mm','> 62 mm')))

# Number of tumors

filtered\_data <- filtered\_data %>%

  mutate(Number\_of\_tumors = as.numeric(Number\_of\_tumors)) %>%

  mutate(Number\_of\_tumors = cut(Number\_of\_tumors, breaks=c(-Inf, 1, Inf), labels=c("1", "> 1")))

filtered\_data <- filtered\_data %>%

  mutate(Surgery = case\_when(

    Surgery %in% c("00") ~ "No",

    Surgery %in% c("10", "12", "13", "14", "15", "16", "17") ~ "Local tumor destruction",

    Surgery %in% c("20", "21", "22", "23", "24", "25", "26") ~ "Wedge or segmental resection",

    Surgery %in% c("30", "36", "37", "38", "50", "51", "52", "59", "60", "61", "66", "75") ~ "Lobectomy"))

filtered\_data$Radiotherapy <- ifelse(filtered\_data$Radiotherapy == "No radiation and/or cancer-directed surgery", "No", "Yes")

filtered\_data <-filtered\_data %>% mutate(Survival\_months = as.numeric(Survival\_months))

filtered\_data$Status <- as.numeric(factor(filtered\_data$Status, levels = c('Alive', 'Dead'))) - 1

filtered\_data <- filtered\_data %>%

mutate\_if(is.character, as.factor)

filtered\_data <- filtered\_data %>%

mutate(Histological\_type = factor(Histological\_type))

plot\_filtered\_distribution <- function(filtered\_data\_) {

plot\_list <- list()

for (colname in colnames(filtered\_data\_)) {

p <- ggplot(filtered\_data\_, aes(x = .data[[colname]])) + geom\_bar() +

scale\_x\_discrete(labels = function(x) str\_wrap(x, width = 10)) +

theme(axis.text.x = element\_text(angle = 40, hjust = 1, size = 5))

plot\_list <- append(plot\_list, list(p))

}

grid.arrange(grobs = plot\_list, ncol = 4)

}

plot\_filtered\_distribution(filtered\_data)

data <- filtered\_data %>%

  data.frame(check.names = T)

#### Overall distrubution ####

overall <- tbl\_summary(filtered\_data,statistic = list(all\_continuous() ~ "{mean} ({sd})"))

data <- filtered\_data %>%

  data.frame(check.names = T)

# unicox

 unicox <- filtered\_data %>% tbl\_uvregression(

    method = coxph,

y = Surv(Survival\_months,Status),

    exponentiate = TRUE,

    pvalue\_fun = function(x) style\_pvalue(x, digits = 2),

    hide\_n = T

  ) %>%

  add\_global\_p() %>%

  bold\_p(t = 0.050) %>%

  bold\_labels() %>%

  italicize\_levels()

##### Multivariate Cox regression analysis ####

fml = as.formula(Surv(Survival\_months, Status) ~ .)

multi\_cox <- coxph(fml, data = filtered\_data) %>%

    tbl\_regression(exponentiate = TRUE,

                   pvalue\_fun = ~style\_pvalue(.x, digits = 2),

                   hide\_n = TRUE,

                   tidy\_fun = broom.helpers::tidy\_parameters) %>%

    add\_global\_p() %>%

    bold\_p(t = 0.050) %>%

    bold\_labels() %>%

    italicize\_levels()

#### Merge tables ####

merged\_table <- tbl\_merge(

tbls = list(overall, unicox, multi\_cox),

tab\_spanner = c("\*\*Overall\*\*", "\*\*Univariate Cox\*\*", "\*\*Multivariate Cox\*\*")

) %>%

bold\_labels() %>%

italicize\_levels() %>%

as\_flex\_table() %>%

autofit() %>%

bold(bold = TRUE, part = "header") %>%

flextable::save\_as\_docx(path = 'Desktop/sseer/r/Merge\_Unicox\_Multicox.docx')

#### Feature Selecting -- Correlation plot ------------------------------------

library(corrplot)

corr\_data <- filtered\_data

corr\_data <- as.data.frame(lapply(corr\_data,as.numeric),check.names = F)

colnames(corr\_data) <- lapply(colnames(corr\_data), function(x){gsub('\\.',' ',x)})

corMatMy <- cor(corr\_data%>%.[,setdiff(names(.),c('Status','Survival\_months'))])

col <- colorRampPalette(c("darkorange", "white", "steelblue"))(20)

corrplot(corMatMy, type = "upper", order = "hclust", col = col,tl.col="black", tl.cex=0.8, tl.srt=70)

tiff(paste('Figures','correation plot.tiff',sep = '/'),width=2000,height = 2000,res=300)

corrplot(corMatMy, type = "upper", order = "hclust", col = col,tl.col="black", tl.cex=0.8, tl.srt=70)

dir.create('Figures')

tiff(paste('Figures', 'correation\_plot.tiff', sep = '/'), width=2000, height = 2000, res=300)

corrplot(corMatMy, type = "upper", order = "hclust", col = col, tl.col="black", tl.cex=0.8, tl.srt=70)

dev.off()

write.csv(corr\_data, file = "Desktop/seer/r/corrdata.csv")

corr\_data <- corr\_data[, !(names(corr\_data) %in% c("Gender", "Fibrosis\_score", "Number\_of\_tumors", "Radiotherapy"))]

write.csv(corr\_data, file = "Desktop/sseer/r/data.csv", row.names = FALSE)