Assignment1

#Topic : Correlation between APOBEC signatures and stage according to age in nonsmokers

#1. Background - APOBEC APOBEC is a kind of cytidine deaminases. The enzyme APOBEC which is found in cancer cells help to increase resistance to the drugs by rapidly changing cancer cells genetically. APOBEC also helps to protect against viral infections of mammals. When mismanifested, APOBEC enzyme can be a major factor of mutations in many kinds of cancer.(Shi, Ke; et al., Structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B, 2017.)

#2. Load dataset loading necessary “library”s and data from the paper.

```
library(readxl)
library(tidyverse)
```

```
## ── Attaching packages
─────────────────────────────────────── tidyverse 1.3.1
──
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.5     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.4     ✓ stringr 1.4.0
## ✓ readr   2.0.2     ✓ forcats 0.5.1
## ── Conflicts
─────────────────────────────────────────
tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
library(gridExtra)
```

```
##
```

```
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine
library(pander)
#readxl::excel_sheets('Chen2020/1-s2.0-
S0092867420307431-mmc1.xlsx')
#readxl::excel_sheets('Chen2020/1-s2.0-
S0092867420307431-mmc2.xlsx')
#readxl::excel_sheets('Chen2020/1-s2.0-
S0092867420307431-mmc4.xlsx')
#readxl::excel_sheets('Chen2020/1-s2.0-
S0092867420307431-mmc5.xlsx')
#readxl::excel_sheets('Chen2020/1-s2.0-
S0092867420307431-mmc6.xlsx')
#d1_1 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc1.xlsx', sheet=1, na ="NA")
d1_2 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc1.xlsx', sheet=2, na ="NA")
#d1_3 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc1.xlsx', sheet=3, na ="NA")
#d1_4 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc1.xlsx', sheet=4, na ="NA")
#d1_5 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc1.xlsx', sheet=5, na ="NA")
#d1_6 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc1.xlsx', sheet=6, na ="NA")
#d1_7 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc1.xlsx', sheet=7, na ="NA")
#d1_8 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc1.xlsx', sheet=8, na ="NA")
#d1_9 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc1.xlsx', sheet=9, na ="NA")
#d1_10 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc1.xlsx', sheet=10, na ="NA")
#d1_11 = read_excel('Chen2020/1-s2.0-S0092867420307431-
```

```
mmc1.xlsx', sheet=11, na ="NA")
#d2_1 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc2.xlsx', sheet=1, na ="NA")
#d2_2 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc2.xlsx', sheet=2, na ="NA")
#d2_3 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc2.xlsx', sheet=3, na ="NA")
#d2_4 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc2.xlsx', sheet=4, na ="NA")
#d2_5 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc2.xlsx', sheet=5, na ="NA")
#d2_6 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc2.xlsx', sheet=6, na ="NA")
#d2_7 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc2.xlsx', sheet=7, na ="NA")
#d4_1 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc4.xlsx', sheet=1, na ="NA")
#d4_2 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc4.xlsx', sheet=2, na ="NA")
#d4_3 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc4.xlsx', sheet=3, na ="NA")
#d4_4 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc4.xlsx', sheet=4, na ="NA")
#d4_5 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc4.xlsx', sheet=5, na ="NA")
#d4_6 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc4.xlsx', sheet=6, na ="NA")
#d5_1 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc5.xlsx', sheet=1, na ="NA")
#d5_2 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc5.xlsx', sheet=2, na ="NA")
#d5_3 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc5.xlsx', sheet=3, na ="NA")
#d5_4 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc5.xlsx', sheet=4, na ="NA")
#d5_5 = read_excel('Chen2020/1-s2.0-S0092867420307431-
```

```
mmc5.xlsx', sheet=5, na ="NA")
#d6_1 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc6.xlsx', sheet=1, na ="NA")
d6_2 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc6.xlsx', sheet=2, na ="NA")
#d6_3 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc6.xlsx', sheet=3, na ="NA")
#d6_4 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc6.xlsx', sheet=4, na ="NA")
#d6_5 = read_excel('Chen2020/1-s2.0-S0092867420307431-
mmc6.xlsx', sheet=5, na ="NA")
```

As a result of loading the data file, it was determined that “Correlation between APOBEC signatures and stage according to age in nonsmokers” could be recognized using d1_2 data and d6_2. After examining the correlation between the age and tumor stage of lung cancer patients in non-smokers using the data of d1_2, we compare the disease pattern between patients with APOBEC characteristic and those without APOBEC using the data of d6_2.

#3. Correlation between stage and age in nonsmokers

#3.1. Transforming data

Make data into more useful form. In this case, we need informations kind of Change the column name so that no spaces are included. Use “filter” to remain information that contains only non-smokers data.

```
colnames(d1_2)[5]<-"Smoking_Status"
pander(d1_2 %>% as.data.frame() %>%
filter(Smoking_Status == "Nonsmoke"))
```

Table continues below

ID	Proteome_Batch	Gender	Age	Smoking_Status	Histology Type
P002	B01-2	Male	73.78	Nonsmoke	ADC
P004	B01-4	Female	52.98	Nonsmoke	SCC

P006	B02-2	Female	46.86	Nonsmoke	ADC
P007	B02-3	Male	67.41	Nonsmoke	ADC
P009	B03-1	Female	53.8	Nonsmoke	ADC
P010	B03-2	Female	56.48	Nonsmoke	ADC
P011	B03-3	Male	59.02	Nonsmoke	ADC
P012	B03-4	Male	61.86	Nonsmoke	ADC
P013	B04-1	Female	44.91	Nonsmoke	ADC
P015	B04-3	Female	54.2	Nonsmoke	ADC
P016	B04-4	Male	59.83	Nonsmoke	ADC
P017	B05-1	Female	71.3	Nonsmoke	ADC
P018	B05-2	Male	65.64	Nonsmoke	ADC
P019	B05-3	Female	78.73	Nonsmoke	ADC
P020	B05-4	Female	66.45	Nonsmoke	ADC
P021	B06-1	Female	85.86	Nonsmoke	ADC
P022	B06-2	Female	63.89	Nonsmoke	ADC
P023	B06-3	Female	56.58	Nonsmoke	ADC
P025	B07-1	Female	75.53	Nonsmoke	ADC
P026	B07-2	Female	64.87	Nonsmoke	ADC
P027	B07-3	Female	70.21	Nonsmoke	ADC
P028	B07-4	Female	50.58	Nonsmoke	ADC
P030	B08-2	Female	49.41	Nonsmoke	ADC
P032	B08-4	Male	78.43	Nonsmoke	ADC
P033	B09-1	Female	51.13	Nonsmoke	ADC
P034	B09-2	Female	63.49	Nonsmoke	ADC
P036	B09-4	Female	61.65	Nonsmoke	ADC
P037	B10-1	Male	61.57	Nonsmoke	ADC
P038	B10-2	Male	59.61	Nonsmoke	ADC
P040	B10-4	Female	47.57	Nonsmoke	ADC
P041	B11-1	Male	40.23	Nonsmoke	SCC
P043	B11-3	Female	78.45	Nonsmoke	ADC
P044	B11-4	Female	72.99	Nonsmoke	ADC
P045	B12-1	Female	76.45	Nonsmoke	ADC

P047	B12-3	Female	60.73	Nonsmoke	ASCC
P048	B12-4	Female	54.74	Nonsmoke	ADC
P049	B13-1	Male	52.96	Nonsmoke	ADC
P051	B13-3	Female	65.28	Nonsmoke	ADC
P052	B13-4	Female	65.65	Nonsmoke	ADC
P054	B14-2	Male	58.49	Nonsmoke	ADC
P055	B14-3	Female	66.65	Nonsmoke	ADC
P056	B14-4	Female	79.61	Nonsmoke	ADC
P057	B15-1	Female	57.78	Nonsmoke	ADC
P058	B15-2	Male	85.66	Nonsmoke	ADC
P059	B15-3	Female	57.44	Nonsmoke	ADC
P060	B15-4	Male	60.6	Nonsmoke	ADC
P062	B16-2	Female	59.93	Nonsmoke	ADC
P063	B16-3	Male	68.56	Nonsmoke	ADC
P067	B17-3	Male	41.88	Nonsmoke	ADC
P068	B17-4	Male	68.07	Nonsmoke	ADC
P069	B18-1	Female	59.95	Nonsmoke	ADC
P070	B18-2	Male	84.11	Nonsmoke	ADC
P071	B18-3	Male	68.61	Nonsmoke	ADC
P072	B18-4	Female	65.45	Nonsmoke	ADC
P074	B19-2	Female	80.16	Nonsmoke	ADC
P075	B19-3	Female	47.69	Nonsmoke	ADC
P076	B19-4	Female	60.9	Nonsmoke	ADC
P077	B20-1	Female	44.31	Nonsmoke	ADC
P079	B20-3	Female	58.4	Nonsmoke	Others
P080	B20-4	Female	75.97	Nonsmoke	ADC
P081	B21-1	Female	60.77	Nonsmoke	ADC
P082	B21-2	Female	69.15	Nonsmoke	ADC
P084	B21-4	Female	58.47	Nonsmoke	ASCC
P085	B22-1	Female	63.58	Nonsmoke	ADC
P086	B22-2	Male	70.3	Nonsmoke	ADC
P087	B22-3	Female	73.26	Nonsmoke	Others

P089	B23-1	Female	74.79	Nonsmoke	ADC
P090	B23-2	Male	65.28	Nonsmoke	ADC
P091	B23-3	Female	64.17	Nonsmoke	ADC
P092	B23-4	Female	59.33	Nonsmoke	ADC
P093	B24-1	Female	73.07	Nonsmoke	ADC
P094	B24-2	Male	50.3	Nonsmoke	ADC
P095	B24-3	Female	66.66	Nonsmoke	ADC
P097	B25-1	Female	66.58	Nonsmoke	ADC
P098	B25-2	Male	46.5	Nonsmoke	ADC
P099	B25-3	Female	60.02	Nonsmoke	ADC
P100	B25-4	Female	76.32	Nonsmoke	ADC
P101	B26-1	Female	55	Nonsmoke	ADC
P102	B26-2	Male	50.84	Nonsmoke	ADC
P103	B26-3	Female	60.21	Nonsmoke	ADC
P104	B26-4	Female	52.82	Nonsmoke	ADC
P109	B28-1	Female	43.36	Nonsmoke	ADC
P110	B28-2	Female	48.2	Nonsmoke	ADC
P111	B28-3	Female	65.68	Nonsmoke	ADC
P112	B28-4	Male	59.01	Nonsmoke	ADC

Stage	EGFR_Status	Primary Tumor Location
IB	others	LUL
IA	exon19del	RLL
IB	WT	RLL
IIA	WT	RLL
IIA	L858R	LLL
IB	exon19del	LUL
IA	exon19del	RLL
IB	exon19del	LLL
IA	WT	RML
IA	exon19del	RLL
IA	exon19del	RUL
IIIA	L858R	RML

IB	L858R	RUL
IA	L858R	LLL
IIB	L858R	LLL
IA	L858R	LUL
IA	L858R	RUL
IB	exon19del	RLL
IB	L858R	RML
IA	WT	RUL
IA	exon19del	RUL
IB	others	LUL
IA	L858R	LLL
IA	exon19del	RLL
IA	exon19del	RLL
IA	L858R	RUL
IA	L858R	RUL
IB	L858R	RUL
IA	exon19del	RLL
IB	WT	LLL
IIIA	WT	NA
IIIA	L858R	RUL
IA	exon19del	RUL
IB	L858R	RUL
IB	L858R	LUL
IA	L858R	LLL
IA	L858R	LUL
IA	L858R.exon19del	RLL
IA	exon19del	RUL
IIIA	exon19del	LUL
IA	L858R	LUL
IA	exon19del	RLL
IA	others	RLL
IA	L858R	LLL

IB	L858R	LLL
IA	exon19del	LUL
IB	L858R	LUL
IB	L858R	LUL
IB	WT	LUL
IB	exon19del	RUL
IA	L858R	LLL
IB	exon19del	RUL
IB	exon19del	RLL
IB	L858R	LLL
IA	L858R	RUL
IB	L858R	RUL
IB	L858R	RUL
IIA	others	LLL
IB	WT	RUL
IB	others	LUL
IIA	L858R	RLL
IB	exon19del	LUL
IA	exon19del	RLL
IIIA	others	RLL
IB	L858R	RUL
IB	WT	RLL
IA	L858R	LUL
IA	exon19del	LUL
IA	others	LUL
IA	others	RML
IV	exon19del	LUL
IA	L858R	LLL
IIIA	exon19del	RML
IB	L858R	LLL
IIA	exon19del	RUL
IA	L858R	LUL

IA	exon19del	RLL
IV	L858R	RLL
IIIB	exon19del	RML
IV	WT	RUL
IV	WT	LLL
IA	exon19del	LUL
IA	exon19del	LLL
IB	L858R	RLL
IB	exon19del	RUL

Find the age distribution of the patient by finding the maximum and minimum values of the patient's age.

```
d1_2 %>% as.data.frame() %>% filter(Smoking_Status == "Nonsmoke") %>% arrange(Age) %>% select(Age) %>% max()
```

```
## [1] 85.86448
```

```
d1_2 %>% as.data.frame() %>% filter(Smoking_Status == "Nonsmoke") %>% arrange(Age) %>% select(Age) %>% min()
```

```
## [1] 40.22724
```

Then using "mutate"function, create a new column "Ages" which represents informations organized in 10-year units. Designate this information as d1_2_Ages.

```
d1_2_Ages <- d1_2 %>% as.data.frame() %>% filter(Smoking_Status == "Nonsmoke") %>% mutate(Ages=ifelse(Age<50,"40s",ifelse(Age<60,"50s",ifelse(Age<70,"60s",ifelse(Age<80,"70s",ifelse(Age>=80,"80s",""))))))
```

#3.2. Drawing plot

Finally, draw the graphs that show the relationship between the tumor stage and the patient's age in non-smokers.

```
pander(d1_2_Ages %>% group_by(Stage, Ages) %>% count())
```

Stage	Ages	n
IA	40s	4

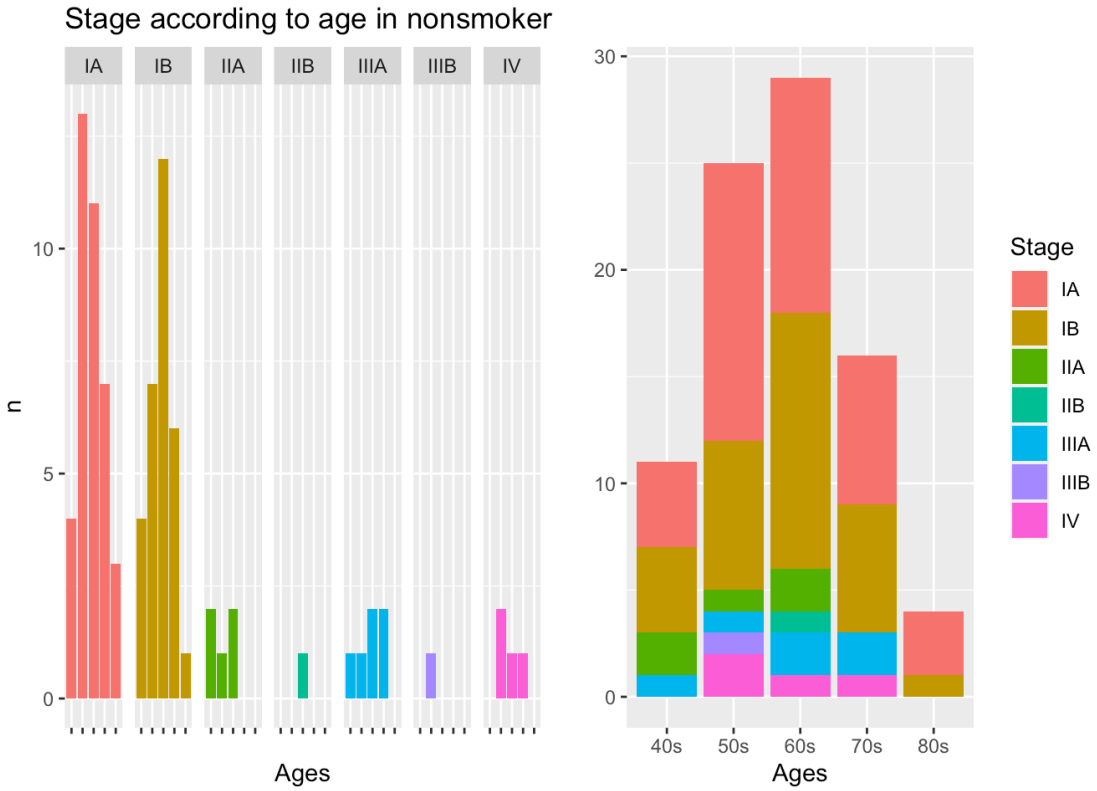
IA	50s	13
IA	60s	11
IA	70s	7
IA	80s	3
IB	40s	4
IB	50s	7
IB	60s	12
IB	70s	6
IB	80s	1
IIA	40s	2
IIA	50s	1
IIA	60s	2
IIB	60s	1
IIIA	40s	1
IIIA	50s	1
IIIA	60s	2
IIIA	70s	2
IIIB	50s	1
IV	50s	2
IV	60s	1
IV	70s	1

```
aa <- d1_2_Ages %>% select(Stage, Ages) %>%
count(Stage, Ages) %>%
  ggplot(aes(Ages, n, fill=Stage)) +
  geom_bar(stat="identity", position="stack") +
  facet_grid(~Stage) +
  theme(legend.position = "none") +
  labs(title = 'Stage according to age in nonsmokers')
+
  scale_x_discrete(label = c("","","","",""))

bb <- d1_2_Ages %>% select(Stage, Ages) %>%
count(Stage, Ages) %>%
  ggplot(aes(Ages, n, fill=Stage)) +
```

```
  geom_bar(stat="identity", position="stack") +
  ylab("") +
  labs(title = '')

grid.arrange(aa, bb, ncol = 2)
```



#3.3. Result and analysis In non-smoker lung cancer patients, stage IA and IB accounted for the most cases and the number of the patients was the smallest in the order of Ages 80s-40s-70s-50s-60s. It is presumed that the tumor stages in all patient Ages have a generally similar pattern. It can be inferred that stage IA and IB do not show specificity according to the age of the patient. The number of patient cases varies according to the Ages of the patient, and it may be thought that the incidence of cancer varies with age. But in my opinion, that thought is an incorrect analysis. We can guess that it is more correct to think this way: the reason why the number of 80s cases is small is because older patients are in risks of death by

tumors.

#4. Correlation between APOBEC signatures and stage according to age in nonsmokers

#4.1. Transforming data

This time, we use data “d6_2. Filter some useful data. Delete unnecessary columns. To utilize the data, change the order of rows and columns of the dataframe and rename the column. Then mutate data to make”Ages" column. Designate this dataframe as “d6_2_data”.

```
d6_2_data <- d6_2[,-c(2,3,4,5)] %>% filter(id %in%
c("Age", "SmokingStatus", "OverallStaging",
"APOBEC_Enriched"))
d6_2_data <- d6_2_data %>% t() %>% as.data.frame()
d6_2_data <- rownames_to_column(d6_2_data, var = "id")
d6_2_data <- rename(d6_2_data,Age = V1,
SmokingStatus=V2, OverallStaging=V3,
APOBEC_Enriched=V4)
d6_2_data <- d6_2_data[-1,]
d6_2_data <- d6_2_data %>% filter(SmokingStatus=="no")
%>%
mutate(Ages=ifelse(Age<50,"40s",ifelse(Age<60,"50s",ife
lse(Age<70,"60s",ifelse(Age<80,"70s",ifelse(Age>=80,"80
s",""))))))
```

#4.2. Drawing plot

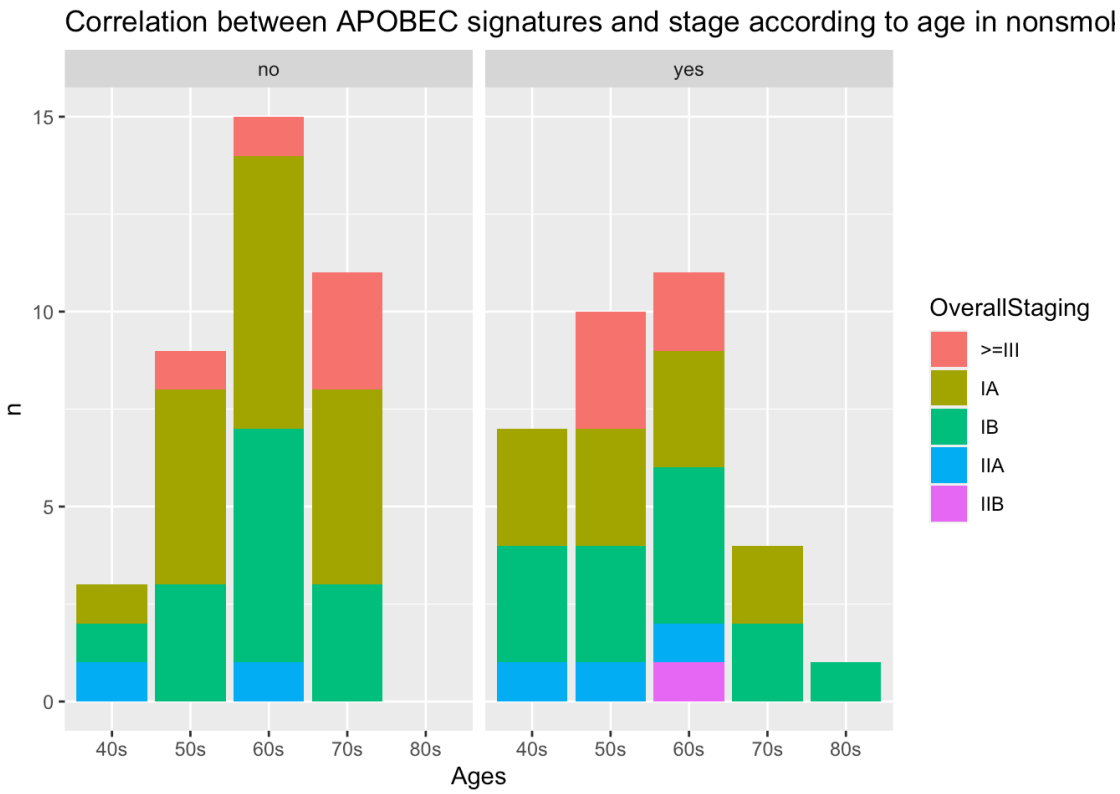
Visualization the graph that shows the relationship between APOBEC signatures and stage according to age in nonsmokers.

```
pander(d6_2_data %>% group_by(APOBEC_Enriched) %>%
count(OverallStaging, Ages))
```

APOBEC_Enriched	OverallStaging	Ages	n
no	>=III	50s	1
no	>=III	60s	1

no	>=III	70s	3
no	IA	40s	1
no	IA	50s	5
no	IA	60s	7
no	IA	70s	5
no	IB	40s	1
no	IB	50s	3
no	IB	60s	6
no	IB	70s	3
no	IIA	40s	1
no	IIA	60s	1
yes	>=III	50s	3
yes	>=III	60s	2
yes	IA	40s	3
yes	IA	50s	3
yes	IA	60s	3
yes	IA	70s	2
yes	IB	40s	3
yes	IB	50s	3
yes	IB	60s	4
yes	IB	70s	2
yes	IB	80s	1
yes	IIA	40s	1
yes	IIA	50s	1
yes	IIA	60s	1
yes	IIB	60s	1

```
d6_2_data %>% group_by(APOBEC_Enriched) %>%
count(OverallStaging, Ages) %>%
  ggplot(aes(Ages, n, fill=OverallStaging)) +
  geom_bar(stat="identity", position="stack") +
  facet_grid(~APOBEC_Enriched) +
  labs(title = 'Correlation between APOBEC signatures
and stage according to age in nonsmokers')
```



“yes” means “APOBEC_Enriched”, and “no” means “not APOBEC_Enriched”

#4.3. Result and analysis We can see that there is only a slight difference in the stage of the tumor depending on the abundance of APOBEC. The APOBEC-enriched side had slightly fewer cases than the non-APOBEC-enriched group. Tumors of stage IA were more common in those not rich in APOBE, and other tumor stages showed similar patterns regardless of APOBEC. There was a difference in tumor stage according to the abundance of APOBEC according to age. In the APOBEC-rich group, the incidence rate was relatively high in those in their 40s, and the incidence was relatively low in those in their 60s and 70s.

#5. Conclusion and Suggestion Only weak correlation was found between APOBEC signatures and tumor stage, but no strong correlation was found. Therefore, it is difficult to regard it as a

meaningful relationship. As a result, The correlation between APOBEC enrichment and tumor stage is difficult to recognize, and it was found that the probability of developing IA stage tumor in the APOBEC enrichment situation was slightly higher than in the non-enrichment situation. It is hoped that the correlation between APOBEC and lung cancer can be closely clarified by using criteria that vary according to the degree of enrichment of APOBEC. In addition, it is hoped that it can be visualized using factors related to the tumor stage, the patient’s age, and whether or not to smoke.