220B_Final_Project

Project Goal:

For this project, suppose a survey was conducted within San Francisco to investigate the number of individual consultations with pharmacists.

- Investigate how the number of consultations with a pharmacist is associated with other variables. As part of this investigation, develop a model for pharmacy managers to estimate the expected number of pharmacist consultations by a new customer within a 4 week period, if the pharmacy was provided with values of the other variables for that new customer.
- For the first person in this dataset, use your fitted model to estimate the probability distribution for his/her number of pharmacy consultations within a 4 week period, i.e., estimate the probability that his/her number of pharmacist consultations equals 0,1,2, etc.
- Clearly state and discuss your model(s) and assumptions, and also the limitations of your analysis in the context of this study. Within the discussion section, you may include any questions you would like to ask the people who designed the survey and collected these data

```
pharmacy_data <- read.table("pharmacist.txt",header = T)

#inspect first few rows
head(pharmacy_data)
```

```
##   pc sex  age income lp fp fr ill ad hs ch1 ch2
## 1  0   1 0.19   0.45  1  0  0   0  0  0   0   0
## 2  0   1 0.72   0.25  0  0  1   2 14  5   0   1
## 3  0   0 0.47   1.30  1  0  0   2  0  1   0   1
## 4  0   0 0.27   0.90  1  0  0   0  0  0   0   0
## 5  0   0 0.19   0.15  0  0  0   2  0  5   0   0
## 6  0   0 0.72   0.45  0  0  1   3  0  0   1   0
```

We are trying to predict pc, which is the number of consultations with a pharmacist in the past 4 weeks.

- pc variable represents counts, so we are modeling a Poisson distribution

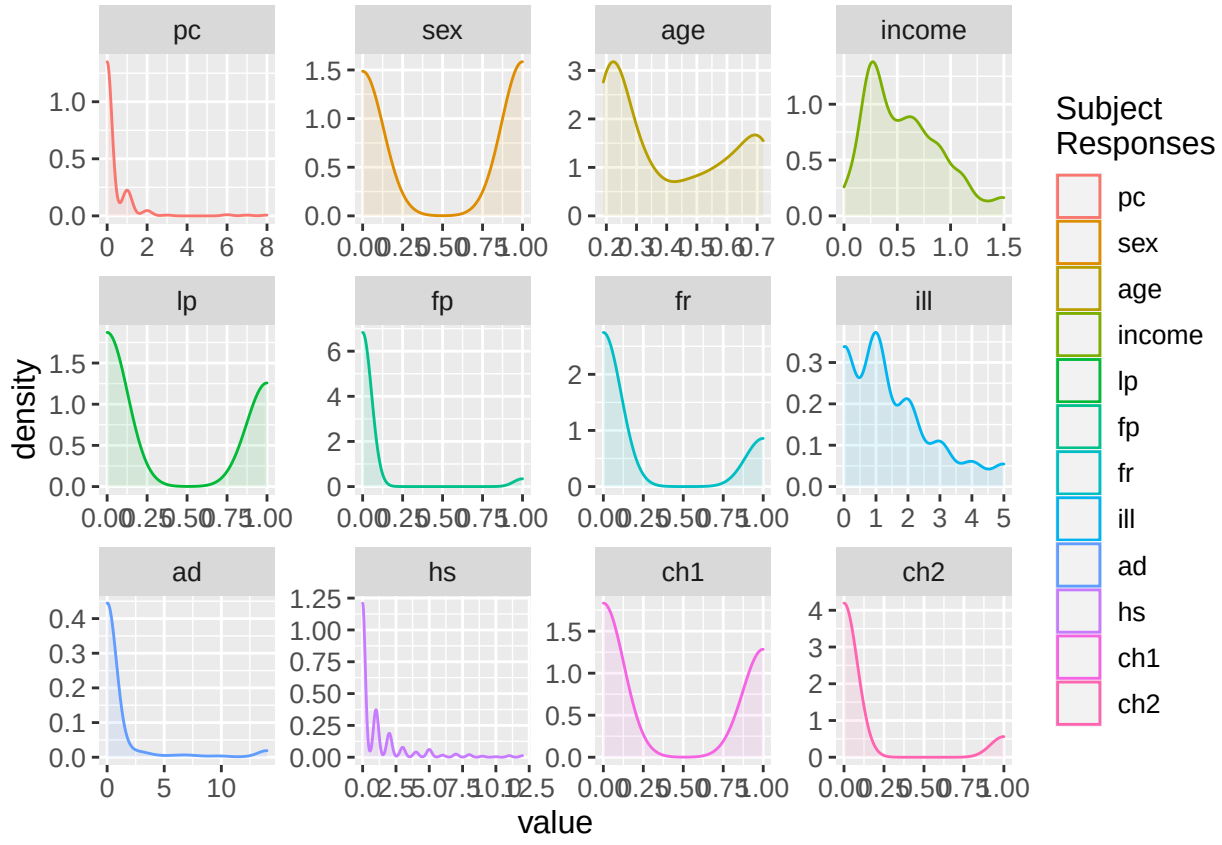
EDA

```
long_df=melt(pharmacy_data)
```

```
## No id variables; using all as measure variables
```

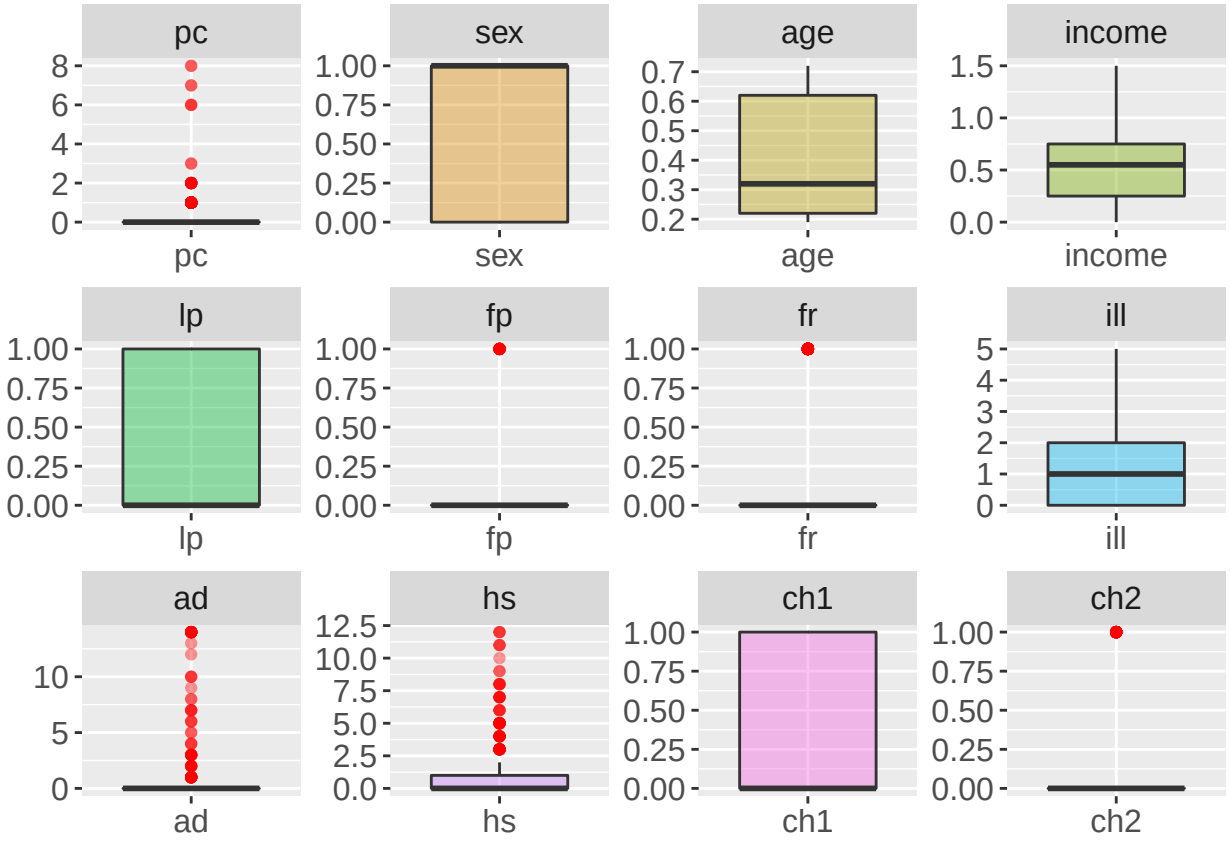
```
kernel_plots <- ggplot(long_df, aes(value, colour = variable)) +
  geom_density(aes(fill=variable),alpha = 0.1)+
  facet_wrap(~variable, scales = "free") +
  labs(color = 'Subject\nResponses') + guides(fill=F) +
  theme(text = element_text(size = 12))

kernel_plots
```



Data either follows a bimodal, exponential, or skewed normal distribution.

```
cov.boxs <- ggplot(stack(pharmacy_data), aes(x = ind, y = values, fill = ind)) +
  geom_boxplot(outlier.color = 'red', alpha = 0.4) + facet_wrap(~ind, scales = 'free')+
  theme(legend.position = "none", axis.title = element_blank(), text = element_text(size = 15))
cov.boxs
```


pc
8
6
4
2
0
pc
sex
1.00
0.75
0.50
0.25
0.00
sex
age
0.7
0.6
0.5
0.4
0.3
0.2
age
income
1.5
1.0
0.5
0.0
income
lp
1.00
0.75
0.50
0.25
0.00
lp
fp
1.00
0.75
0.50
0.25
0.00
fp
fr
1.00
0.75
0.50
0.25
0.00
fr
ill
5
4
3
2
1
0
ill
ad
10
5
0
ad
hs
12.5
10.0
7.5
5.0
2.5
0.0
hs
ch1
1.00
0.75
0.50
0.25
0.00
ch1
ch2
1.00
0.75
0.50
0.25
0.00
ch2