Airbnb Bookings Exploratory Data Analasis

Diana Liang

```
# read in data
dat <- read_csv('bookings_sample.csv')
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   date_account_created = col_date(format = ""),
##   timestamp_first_active = col_double(),
##   date_first_booking = col_date(format = ""),
##   age = col_double(),
##   signup_flow = col_double(),
##   treat = col_double(),
##   is_booked = col_double(),
##   is_eng = col_double(),
##   is_mobile = col_double(),
##   date_first_active = col_date(format = ""),
##   days_diff = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
# check duplicates
n_distinct(dat$id)==nrow(dat)
```

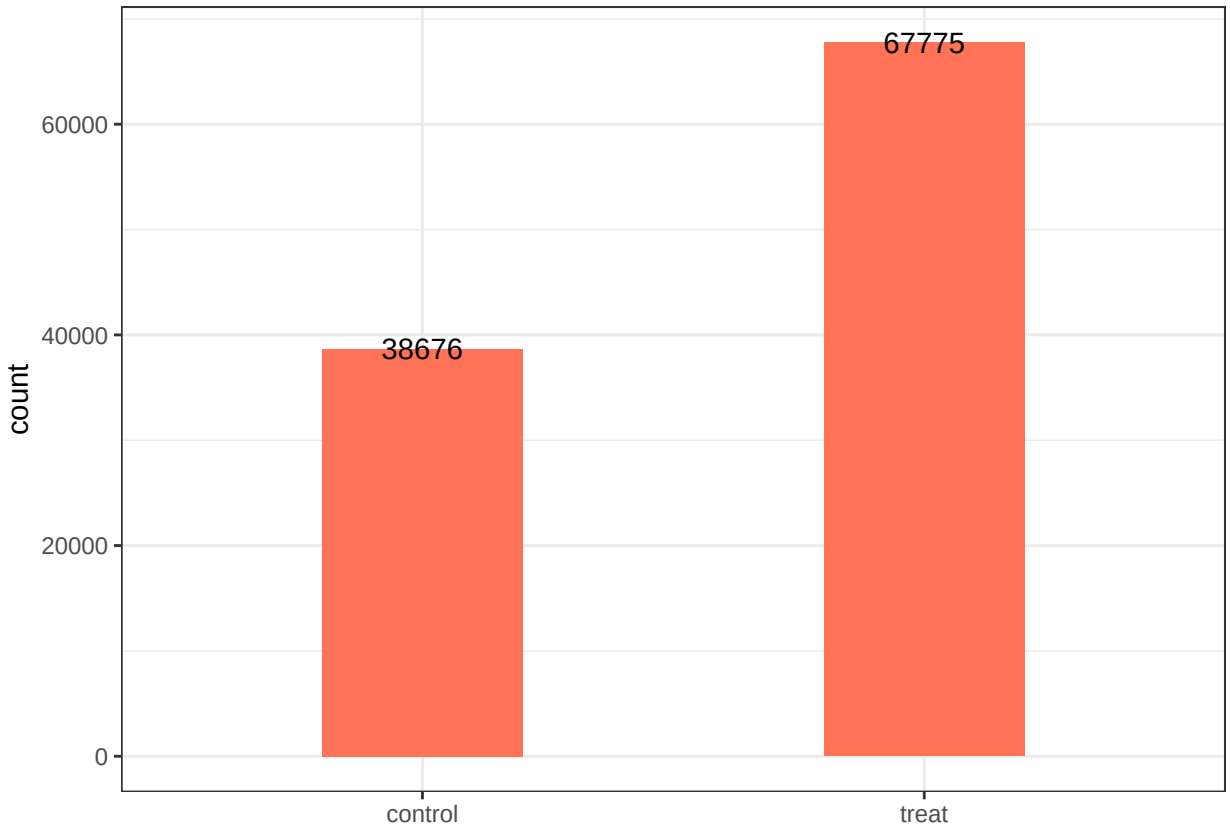
```
## [1] TRUE
```

```
# all first active before sign up?
sum(dat$date_first_active<=dat$date_account_created)==nrow(dat)
```

```
## [1] TRUE
```

Treatment: direct marketing or not?

```
dat %>% group_by(treat) %>%
  summarise(n=n()) %>%
  ggplot(aes(factor(ifelse(treat==1,'treat','control')),n,label=n)) +
  geom_bar(stat='identity',width=.4,fill='coral1') +
  labs(x=element_blank(),y='count') +
  geom_text(position=position_dodge(width=1))
```

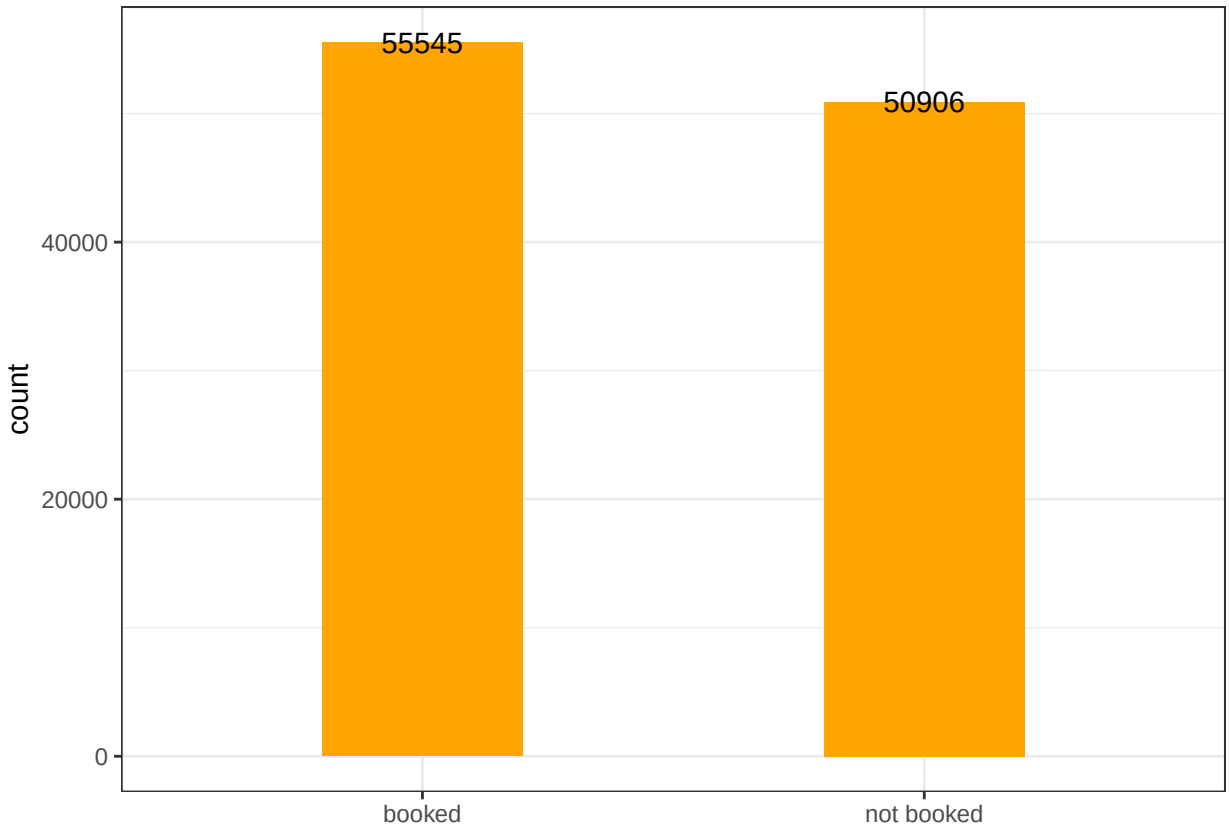


More treat than control.

Outcome: Booked or not?

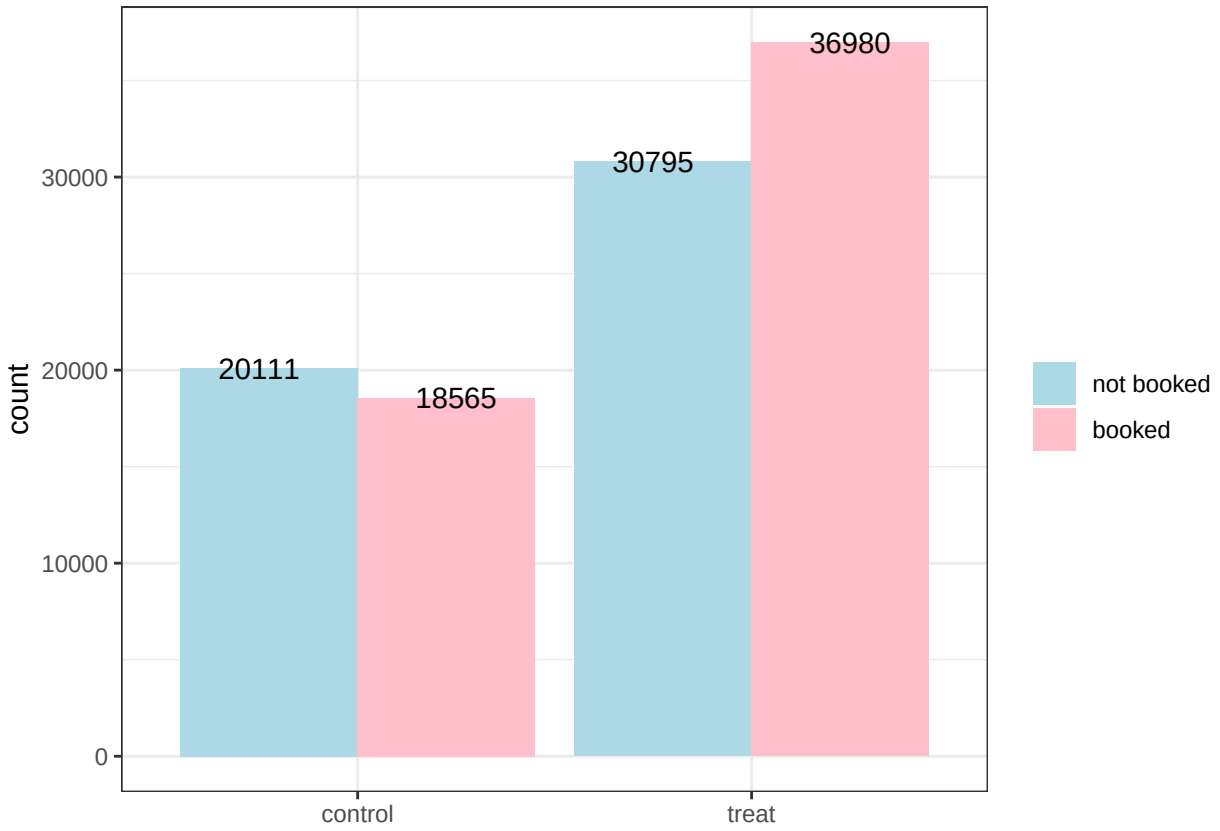
Overall distribution

```
# overall
dat %>% group_by(is_booked) %>%
  summarise(n=n()) %>%
  ggplot(aes(factor(ifelse(is_booked==1,'booked','not booked')),n,label=n)) +
  geom_bar(stat='identity',width=.4,fill='orange') +
  labs(x=element_blank(),y='count') +
  geom_text(position=position_dodge(width=1))
```



Overall, slightly more booked than not booked. ## Distributions within treatment and control

```
# within groups
dat %>% group_by(is_booked,treat) %>%
  summarise(n=n()) %>%
  ggplot(aes(factor(ifelse(treat==1,'treat','control')),n,fill=factor(is_booked),label=n)) +
  geom_bar(stat='identity',position='dodge') +
  theme(legend.title=element_blank()) +
  labs(x=element_blank(),y="count") +
  scale_fill_manual(
    name='Group',
    labels=c('not booked','booked'),
    values=c('lightblue','pink')) +
  geom_text(position=position_dodge(width=1))
```

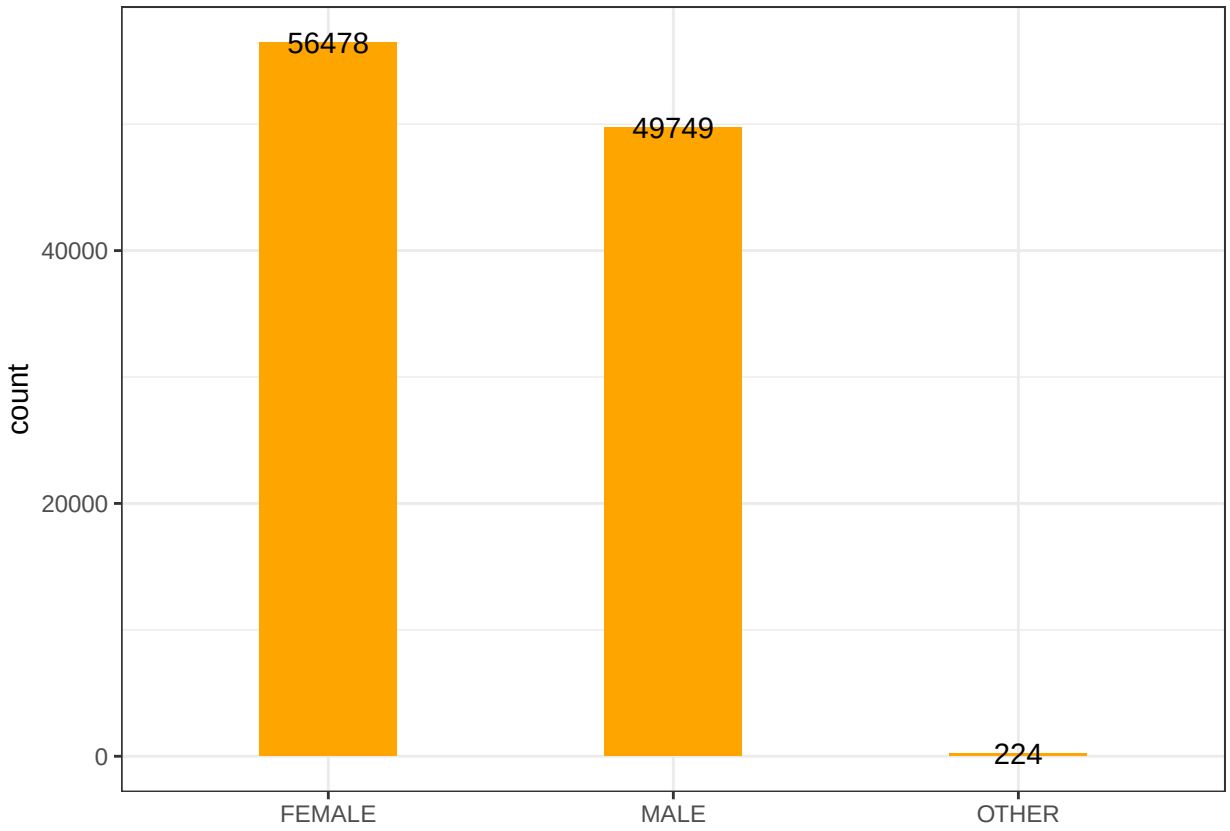


In treatment group, booked more than not booked. In control group, booked slightly less than not booked.

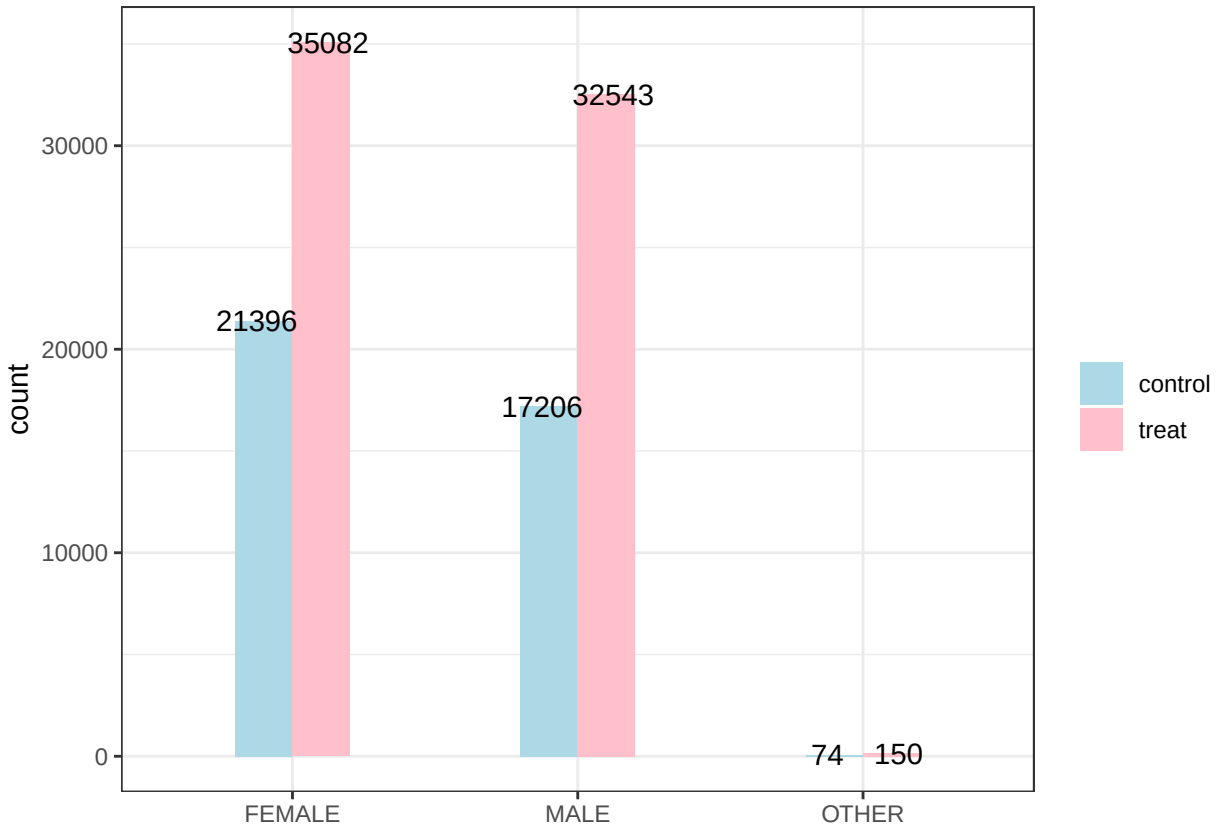
Covariates

Gender

```
# overall
dat %>% group_by(gender) %>%
  summarise(n=n()) %>%
  ggplot(aes(gender,n,label=n)) +
  geom_bar(stat='identity',width=.4,fill='orange') +
  labs(x=element_blank(),y='count') +
  geom_text(position=position_dodge(width=1))
```



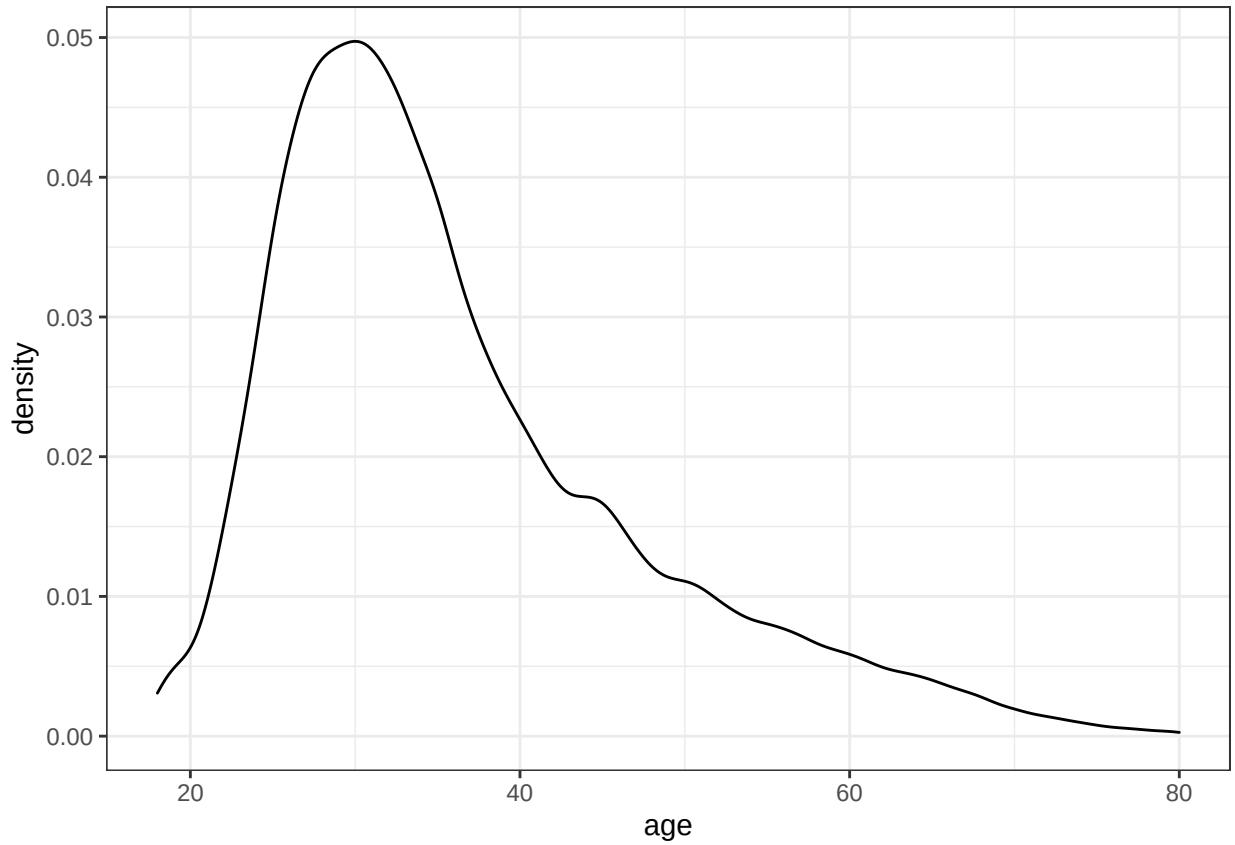

```
# by groups
dat %>% group_by(gender,treat) %>%
  summarise(n=n()) %>%
  ggplot(aes(gender,n,label=n,fill=factor(treat))) +
  geom_bar(stat='identity',position='dodge',width=.4) +
  labs(x=element_blank(),y='count') +
  geom_text(position=position_dodge(width=.5)) +
  theme(legend.title=element_blank()) +
  scale_fill_manual(
    name='Group',
    labels=c('control','treat'),
    values=c('lightblue','pink'))
```



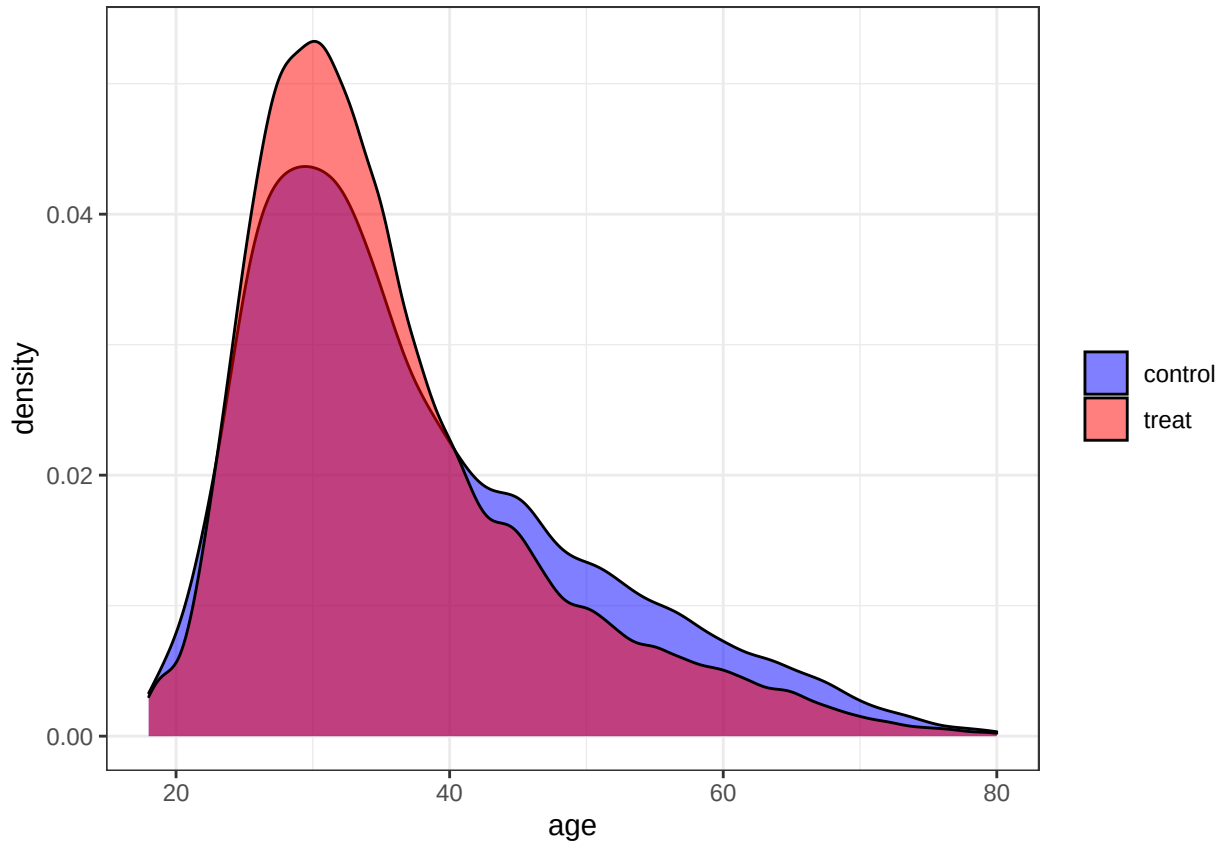
More female than male, very few other (probably to exclude them?). All gender has good overlap. Within each gender, more treat than control.

Age

```
# overall
dat %>% ggplot(aes(age)) +
  geom_density()
```



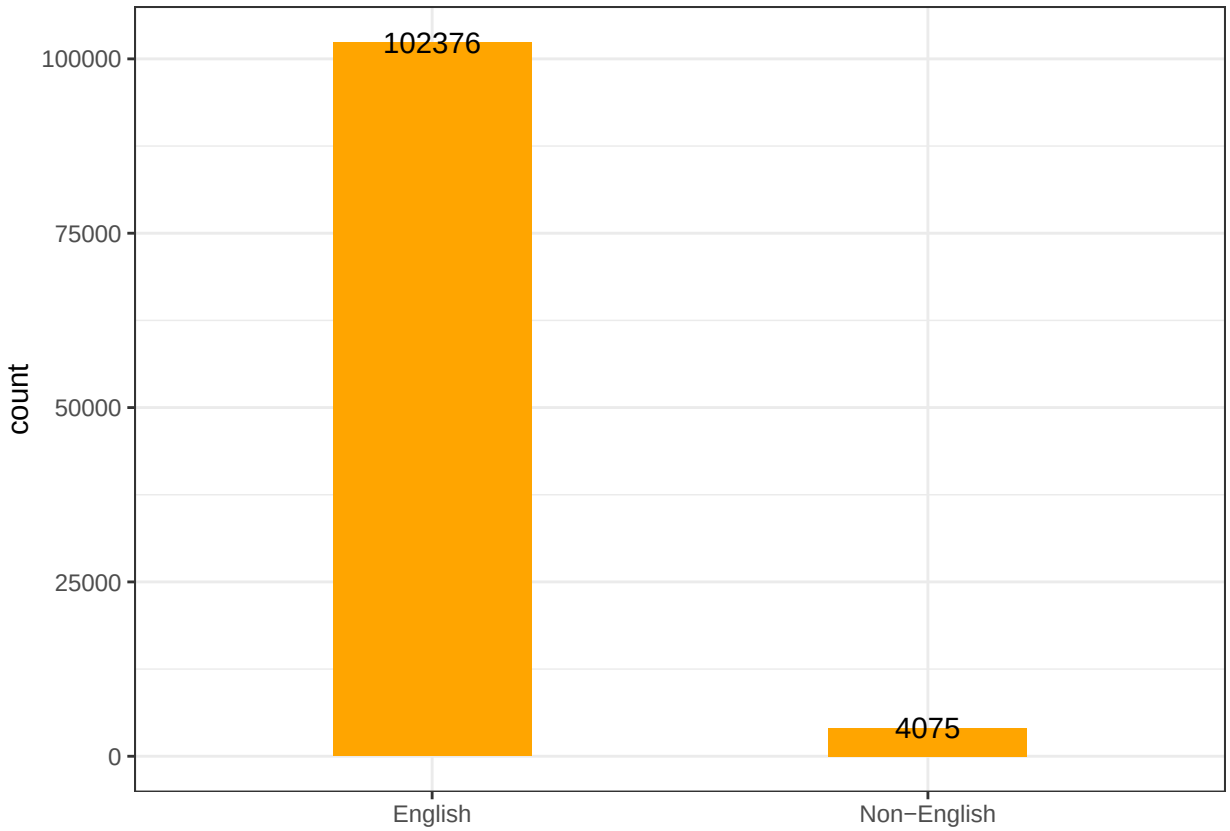

```
# age by group
dat %>% ggplot() +
  geom_density(aes(age,fill=as.factor(treat)),alpha=.5) +
  scale_fill_manual(
    labels=c('control','treat'),
    values=c('blue','red')
  ) +
  theme(legend.title=element_blank())
```



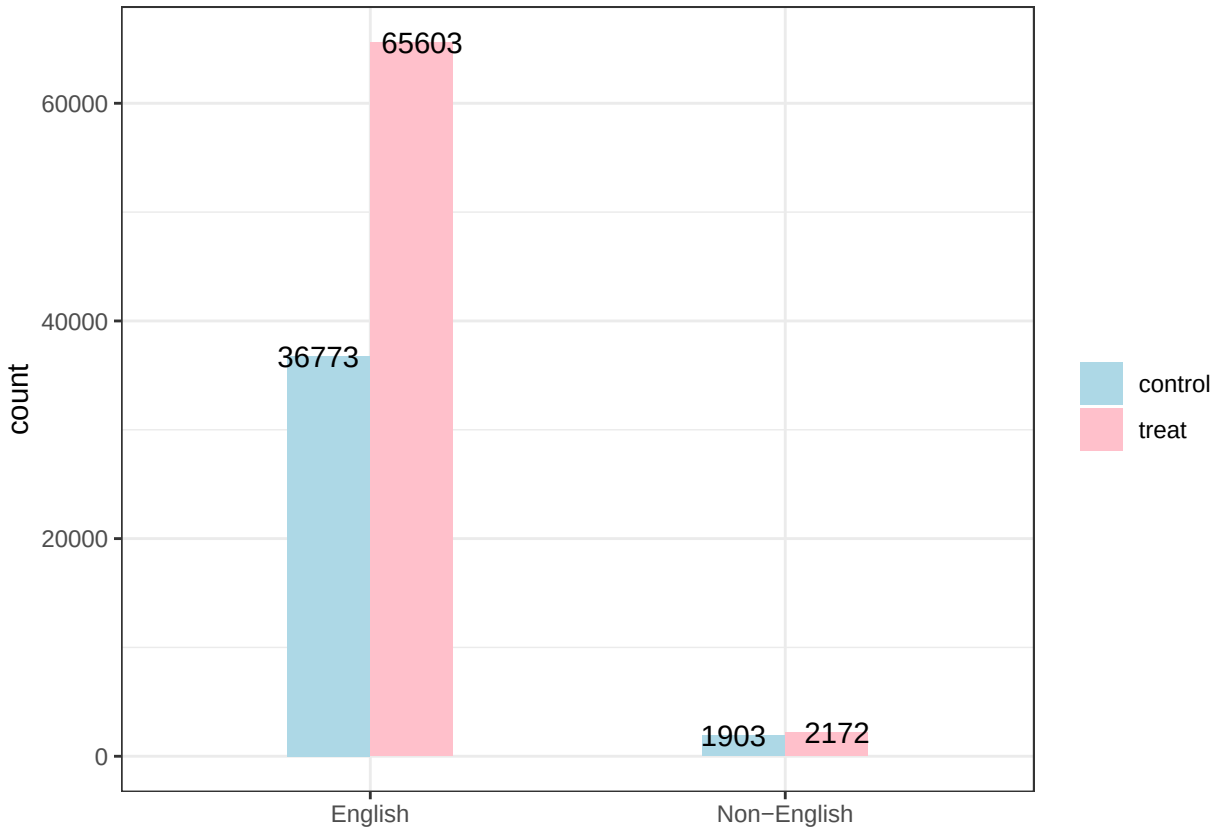
Age is right-skewed. Good overlap: both control and treat cover the whole age range. Balance looks fine for age.

English or not?

```
# overall
dat %>% group_by(is_eng) %>%
  summarise(n=n()) %>%
  ggplot(aes(factor(ifelse(is_eng==1,'English','Non-English')),n,label=n)) +
  geom_bar(stat='identity',width=.4,fill='orange') +
  labs(x=element_blank(),y='count') +
  geom_text(position=position_dodge(width=1))
```



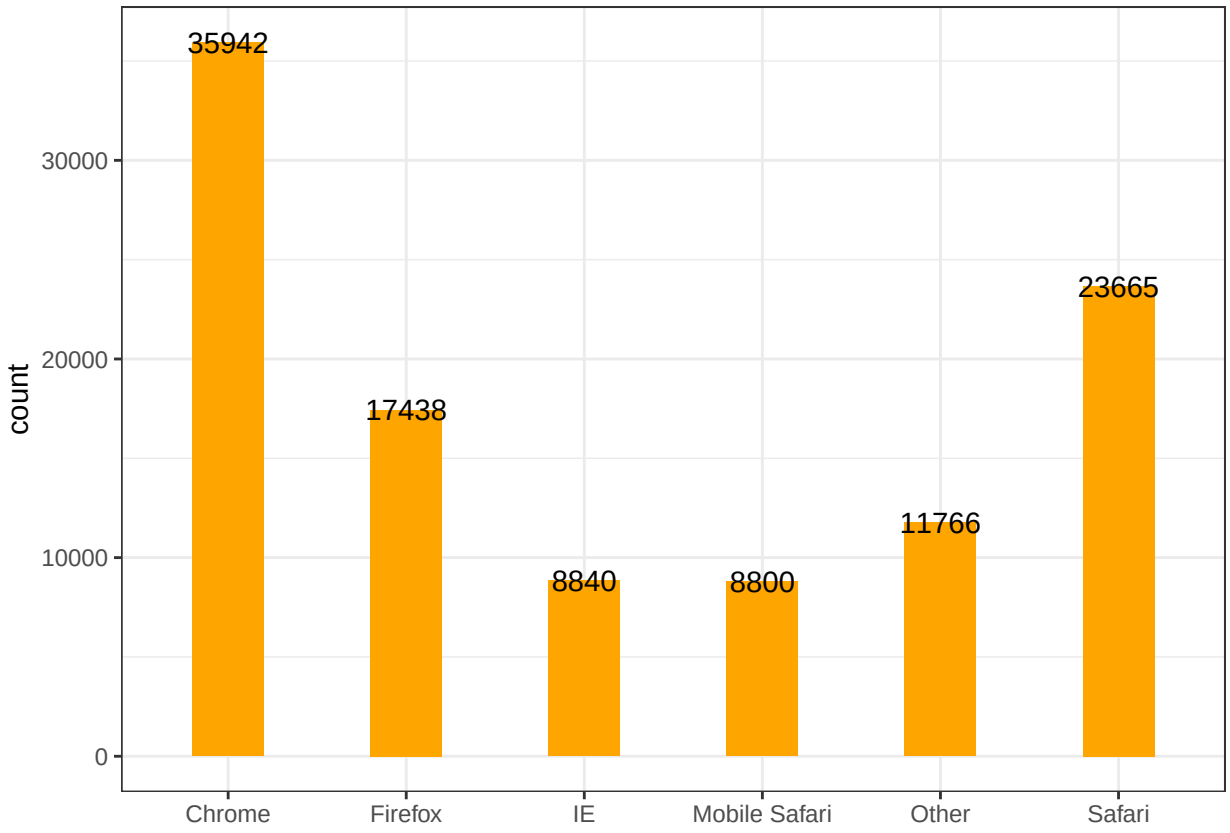

```
# by groups
dat %>% group_by(is_eng,treat) %>%
  summarise(n=n()) %>%
  ggplot(aes(factor(ifelse(is_eng==1,'English','Non-English')),n,label=n,fill=factor(treat))) +
  geom_bar(stat='identity',position='dodge',width=.4) +
  labs(x=element_blank(),y='count') +
  geom_text(position=position_dodge(width=.5)) +
  theme(legend.title=element_blank()) +
  scale_fill_manual(
    name='Group',
    labels=c('control','treat'),
    values=c('lightblue','pink'))
```



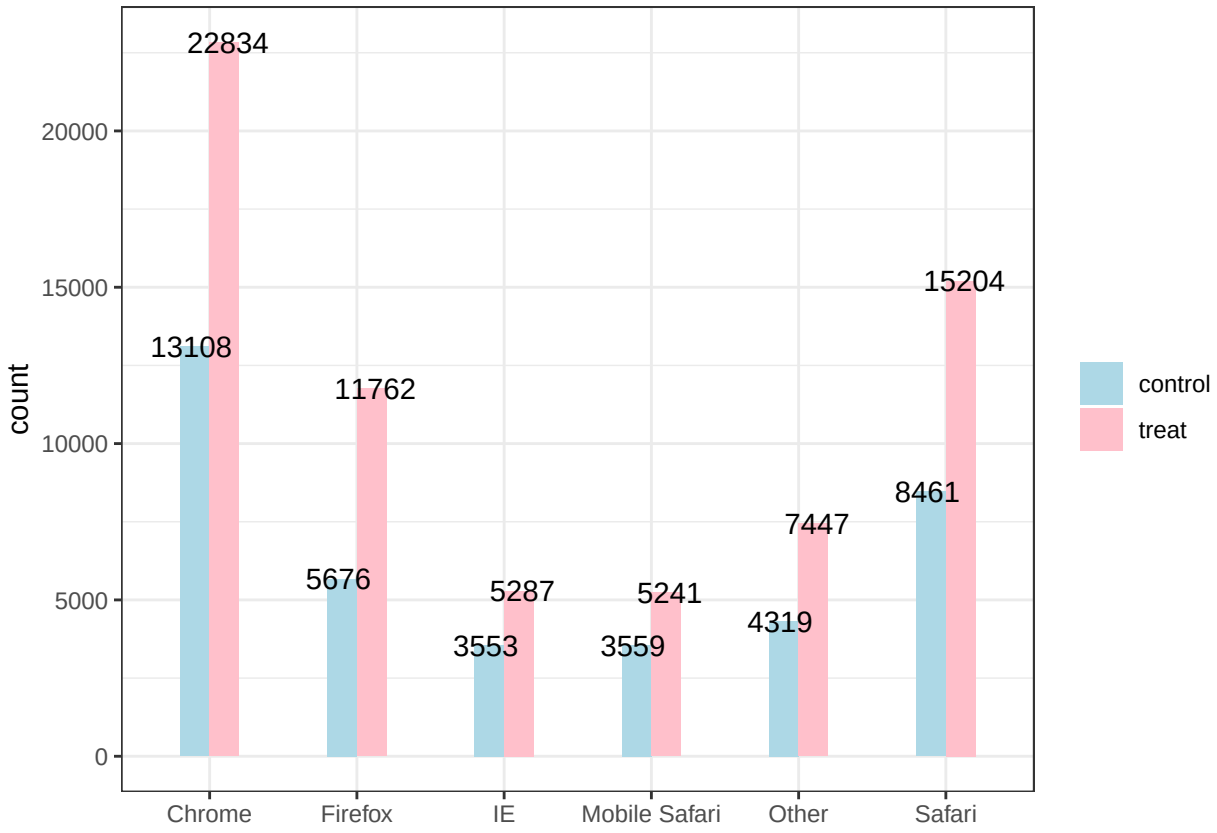
96% use English as language preference on Airbnb. Good overlap: both treatment and control have English and Non-English. Not perfect balance for language preference.

Browser type

```
# overall
dat %>% group_by(browser_group) %>%
  summarise(n=n()) %>%
  ggplot(aes(browser_group,n,label=n)) +
  geom_bar(stat='identity',width=.4,fill='orange') +
  labs(x=element_blank(),y='count') +
  geom_text(position=position_dodge(width=1))
```



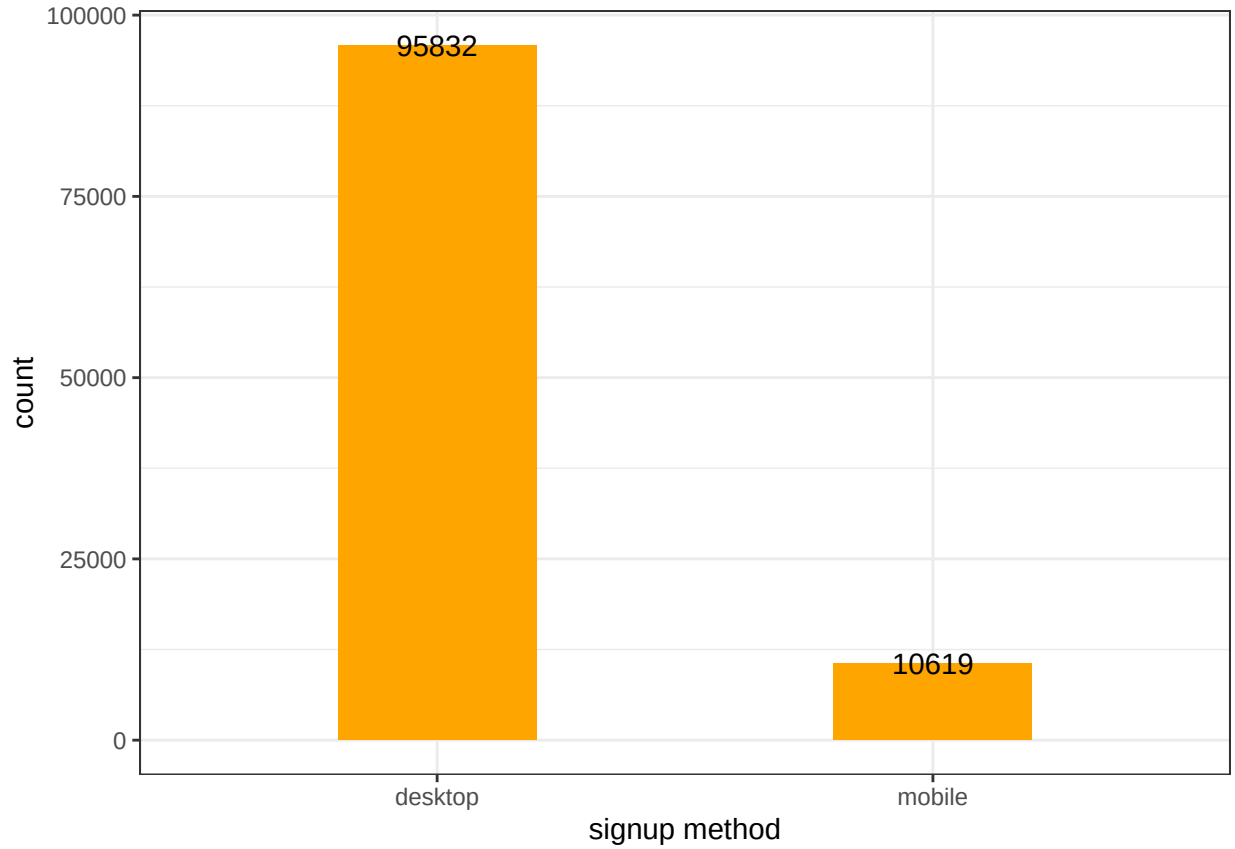

```
# by groups
dat %>% group_by(browser_group,treat) %>%
  summarise(n=n()) %>%
  ggplot(aes(browser_group,n,label=n,fill=factor(treat))) +
  geom_bar(stat='identity',position='dodge',width=.4) +
  labs(x=element_blank(),y='count') +
  geom_text(position=position_dodge(width=.5)) +
  theme(legend.title=element_blank()) +
  scale_fill_manual(
    name='Group',
    labels=c('control','treat'),
    values=c('lightblue','pink'))
```



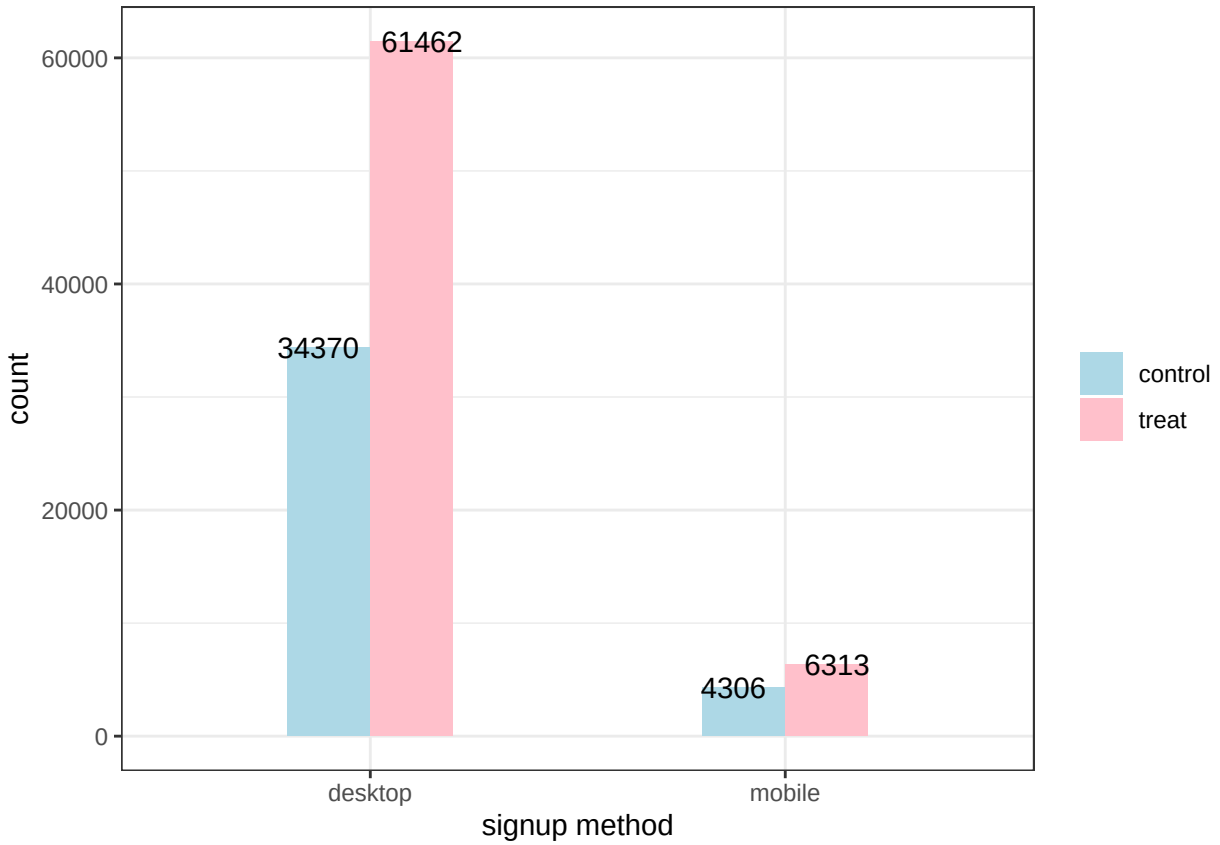
Good overlap and balance for browser type.

Signup method

```
# overall
dat %>% group_by(is_mobile) %>%
  summarise(n=n()) %>%
  ggplot(aes(factor(ifelse(is_mobile==1,'mobile','desktop')),n,label=n)) +
  geom_bar(stat='identity',width=.4,fill='orange') +
  labs(x='signup method',y='count') +
  geom_text(position=position_dodge(width=1))
```



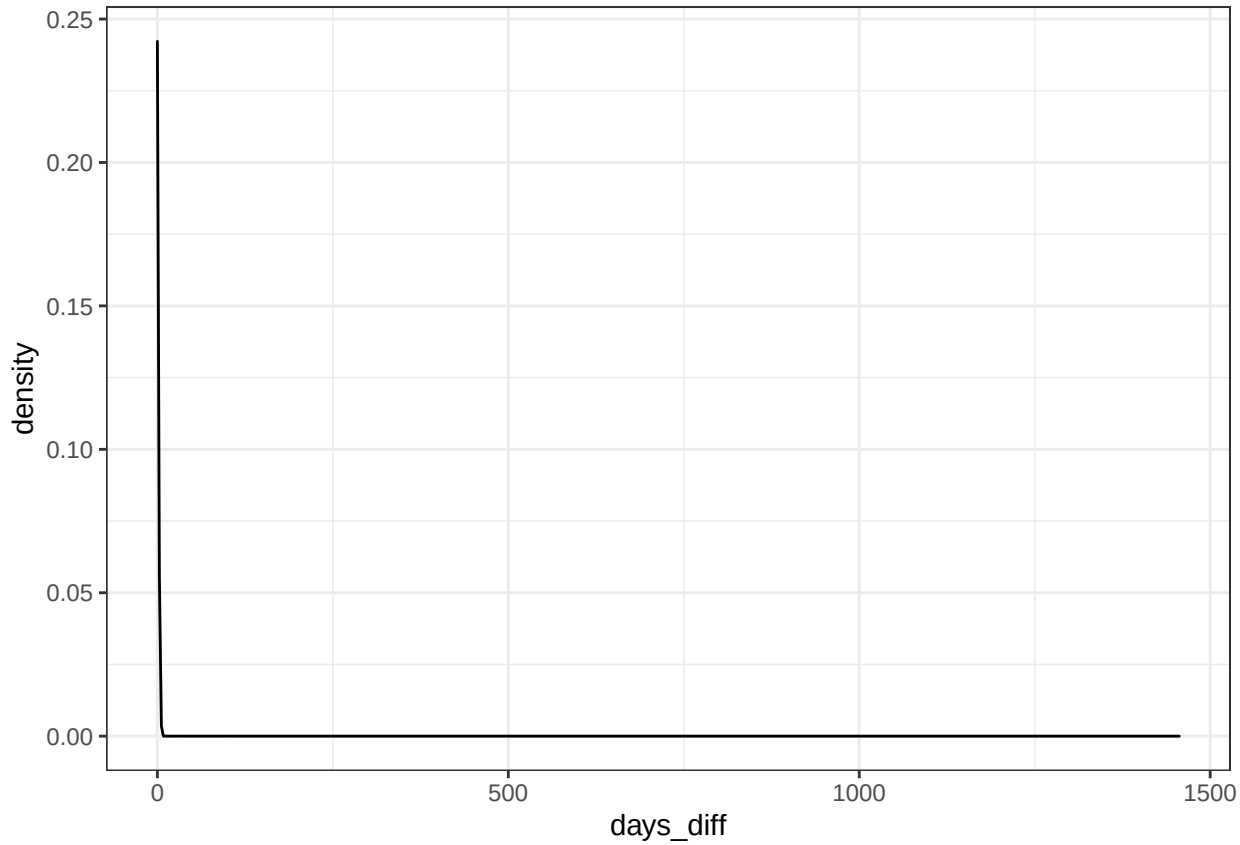

```
# by groups
dat %>% group_by(is_mobile,treat) %>%
  summarise(n=n()) %>%
  ggplot(aes(factor(ifelse(is_mobile==1,'mobile','desktop')),n,label=n,fill=factor(treat))) +
  geom_bar(stat='identity',position='dodge',width=.4) +
  labs(x='signup method',y='count') +
  geom_text(position=position_dodge(width=.5)) +
  theme(legend.title=element_blank()) +
  scale_fill_manual(
    name='Group',
    labels=c('control','treat'),
    values=c('lightblue','pink'))
```



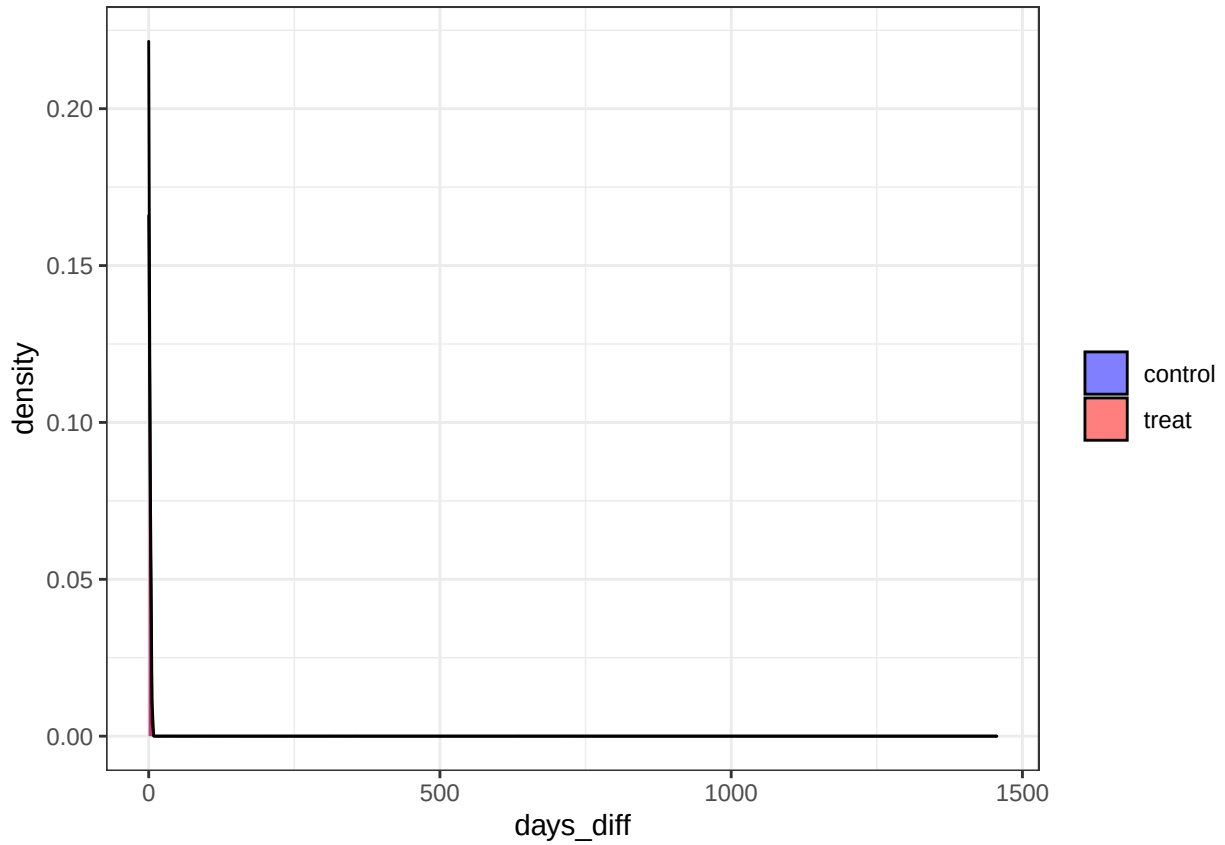
Most of the Airbnb users sign up on desktop. Good overlap and balance for signup method.

Days to signup

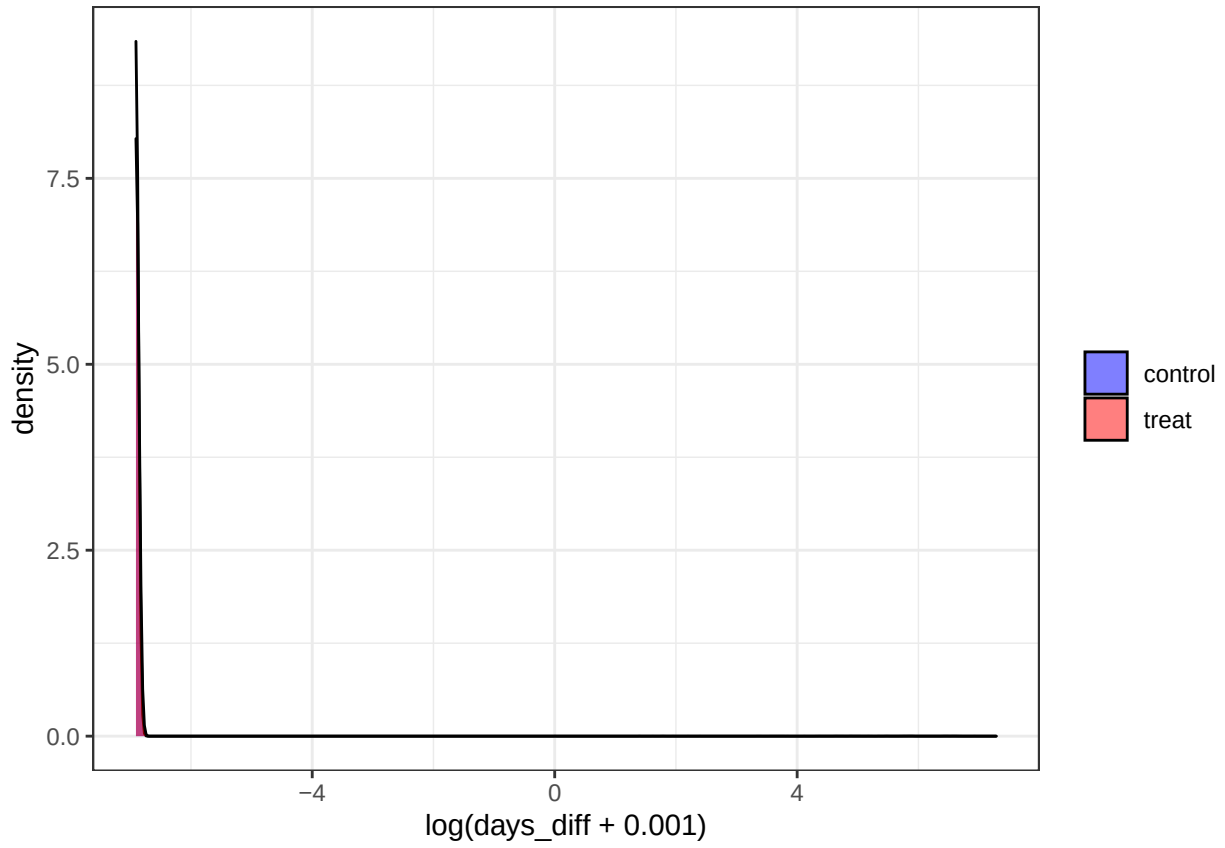
```
# overall
dat %>% ggplot(aes(days_diff)) +
  geom_density()
```




```
# age by group
dat %>% ggplot() +
  geom_density(aes(days_diff,fill=as.factor(treat)),alpha=.5) +
  scale_fill_manual(
    labels=c('control','treat'),
    values=c('blue','red')
  ) +
  theme(legend.title=element_blank())
```




```
# on log scale
dat %>% ggplot() +
  geom_density(aes(log(days_diff+.001),fill=as.factor(treat)),alpha=.5) +
  scale_fill_manual(
    labels=c('control','treat'),
    values=c('blue','red')
  ) +
  theme(legend.title=element_blank())
```

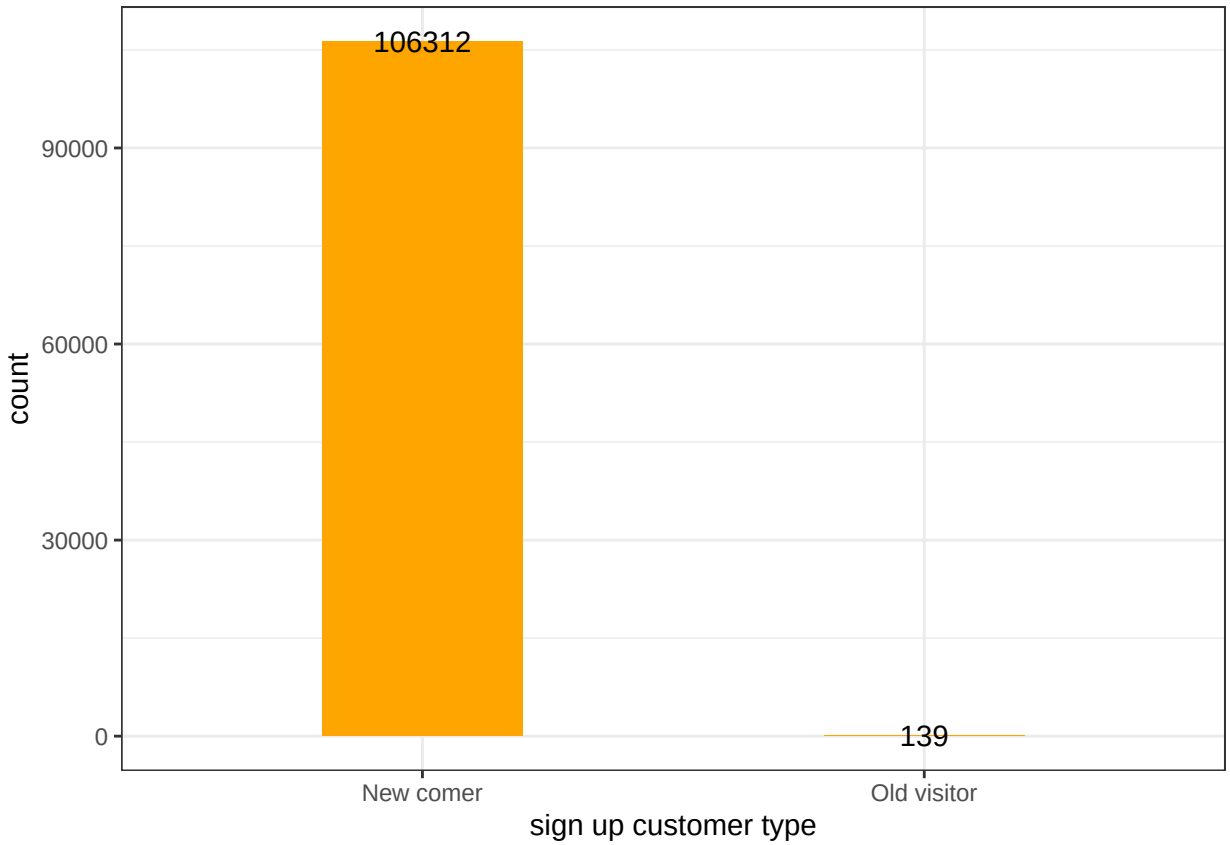


Assumption: marketing happens after signup behavior. Zero-inflated days to signup: hard to see overlap and balance. Might switch to binary indicator for new comer signup.

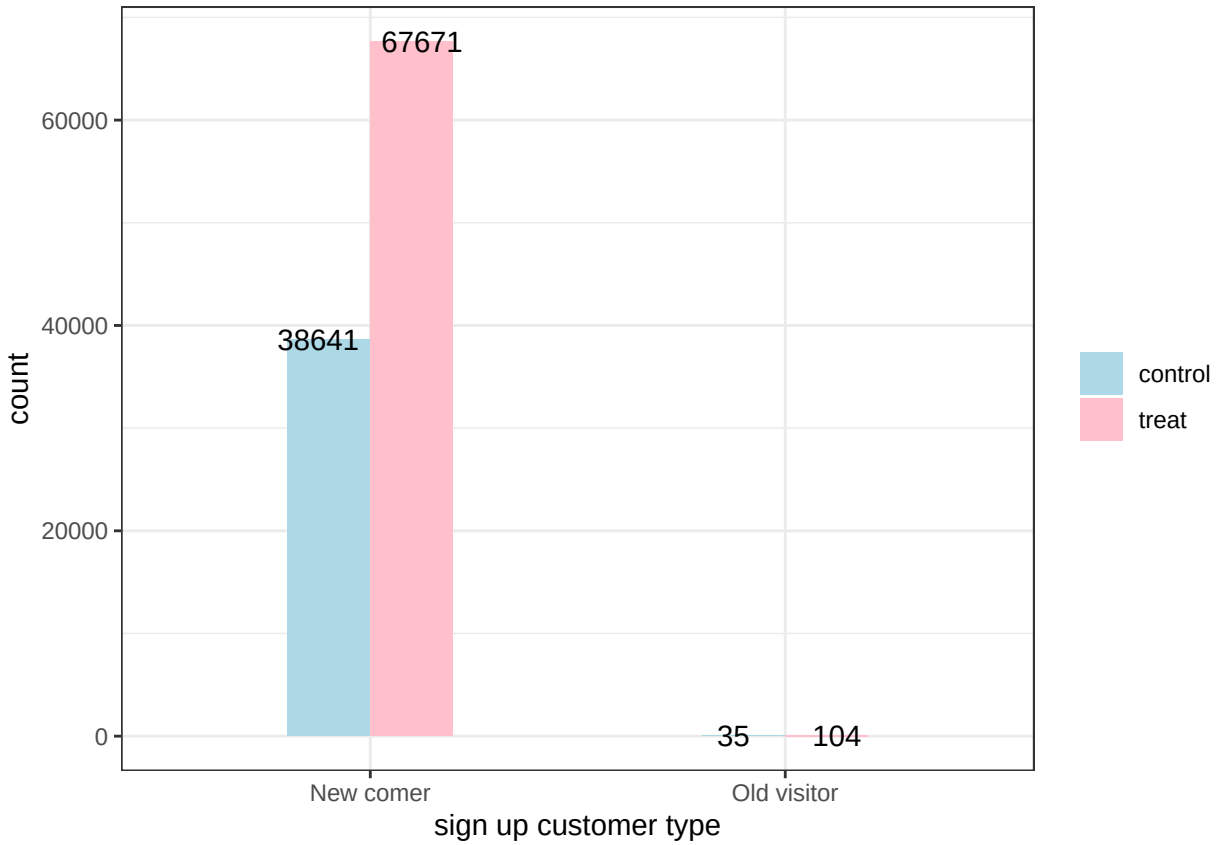
New-comer signup?

```
dat <- dat %>% mutate(new_signup=ifelse(days_diff==0,1,0))

# overall
dat %>% group_by(new_signup) %>%
  summarise(n=n()) %>%
  ggplot(aes(factor(ifelse(new_signup==1,'New comer','Old visitor')),n,label=n)) +
  geom_bar(stat='identity',width=.4,fill='orange') +
  labs(x='sign up customer type',y='count') +
  geom_text(position=position_dodge(width=1))
```




```
# by groups
dat %>% group_by(new_signup,treat) %>%
  summarise(n=n()) %>%
  ggplot(aes(factor(ifelse(new_signup==1,'New comer','Old visitor')),n,label=n,fill=factor(treat))) +
  geom_bar(stat='identity',position='dodge',width=.4) +
  labs(x='sign up customer type',y='count') +
  geom_text(position=position_dodge(width=.5)) +
  theme(legend.title=element_blank()) +
  scale_fill_manual(
    name='Group',
    labels=c('control','treat'),
    values=c('lightblue','pink'))
```



Too imbalanced to be included?