Appendix 2

```
library(ggplot2)
library(summarytools)
library(readr)
library(magrittr)
library(dplyr)
library(viridis)
library(tidyquant)
library(stargazer)
```

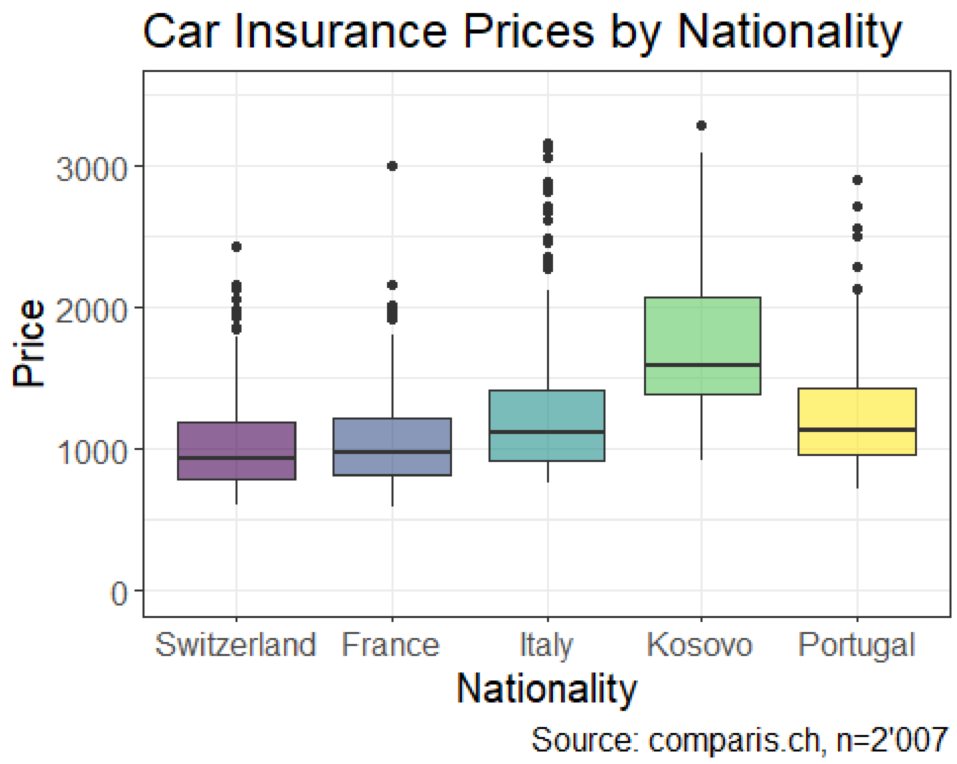
Loading Data and recoding

```
data_insurance <- read_csv('master_df.csv')

data_insurance$gender <- recode(data_insurance$gender, f = "Female", m = "Mal
e")
data_insurance$nationality <- recode(data_insurance$nationality, Schweiz = "S
witzerland",
                                     Frankreich = "France",
                                     Italien = "Italy")
data_insurance$nationality <- relevel(as.factor(data_insurance$nationality),
ref = "Switzerland")
```

Descriptive Analysis

Nationality and Gender plots

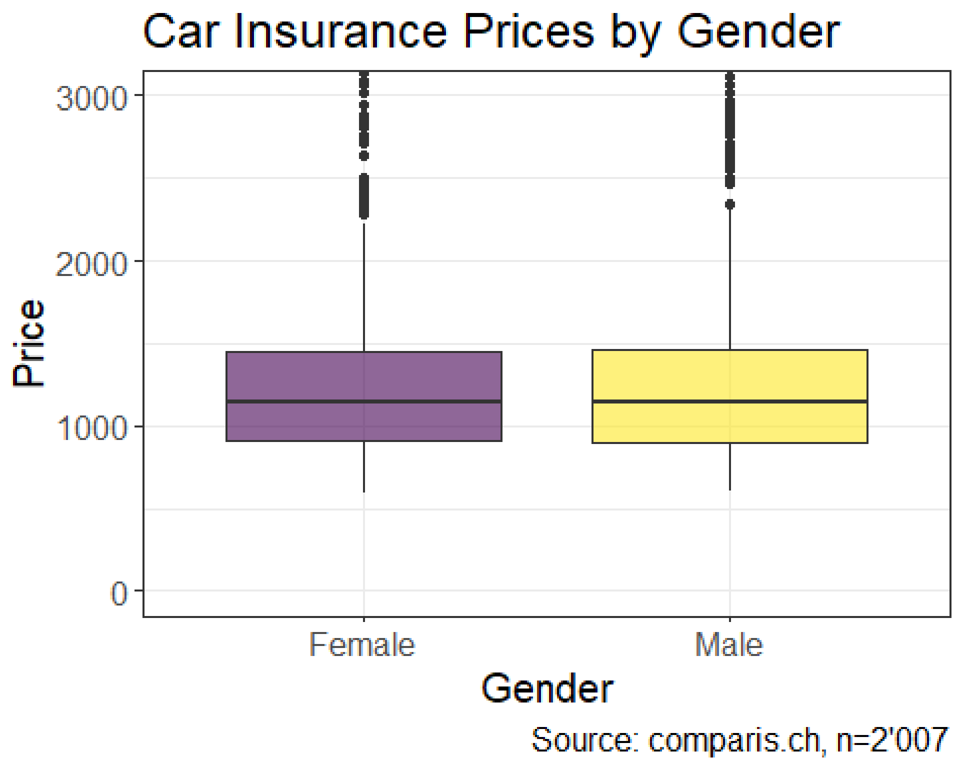
```
boxplot_nationality <- ggplot(data_insurance, aes(x=nationality, y = price, f
ill=nationality))+
  geom_boxplot() +
  scale_fill_viridis(discrete = T, alpha = 0.6)+
        theme_bw()+
        theme(legend.position = 'none')+
  coord_cartesian(ylim = c(0, 3500))+
  labs(x= "Nationality", y="Price",
       title = "Car Insurance Prices by Nationality",
       caption = "Source: comparis.ch, n=2'007")+
  theme(text = element_text(size = 15))
boxplot_nationality
```



Car Insurance Prices by Nationality

Source: comparis.ch, n=2'007

```
median_nationality <- data_insurance %>%
  group_by(nationality) %>%
  summarise(median = median(price))

ggsave("Export/desc_nationality.png", plot = boxplot_nationality, width = 8,
height = 6)

boxplot_gender <- ggplot(data_insurance, aes(x=gender, y = price, fill = gend
er))+
  geom_boxplot() +
  scale_fill_viridis(discrete = T, alpha = 0.6)+
        theme_bw()+
        theme(legend.position = 'none')+
  labs(x= "Gender", y="Price",
       title = "Car Insurance Prices by Gender",
       caption = "Source: comparis.ch, n=2'007")+
  coord_cartesian(ylim= c(0, 3000))+
  theme(text = element_text(size = 15))
boxplot_gender
```



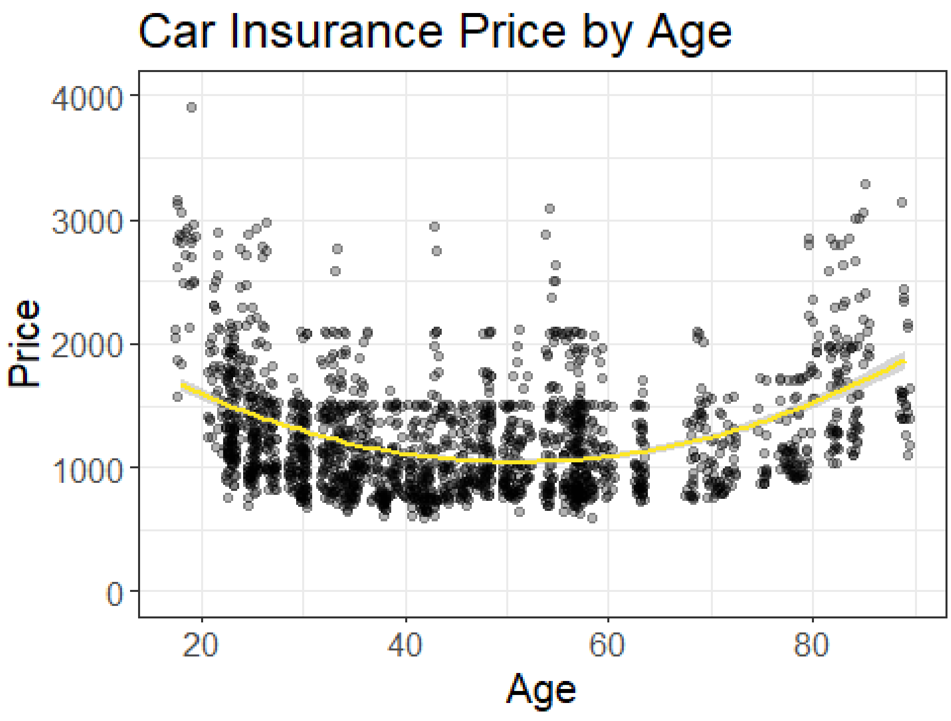
Car Insurance Prices by Gender

Source: comparis.ch, n=2'007

```
ggsave("Export/desc_gender.png", plot = boxplot_gender, width = 8, height = 6
)
```

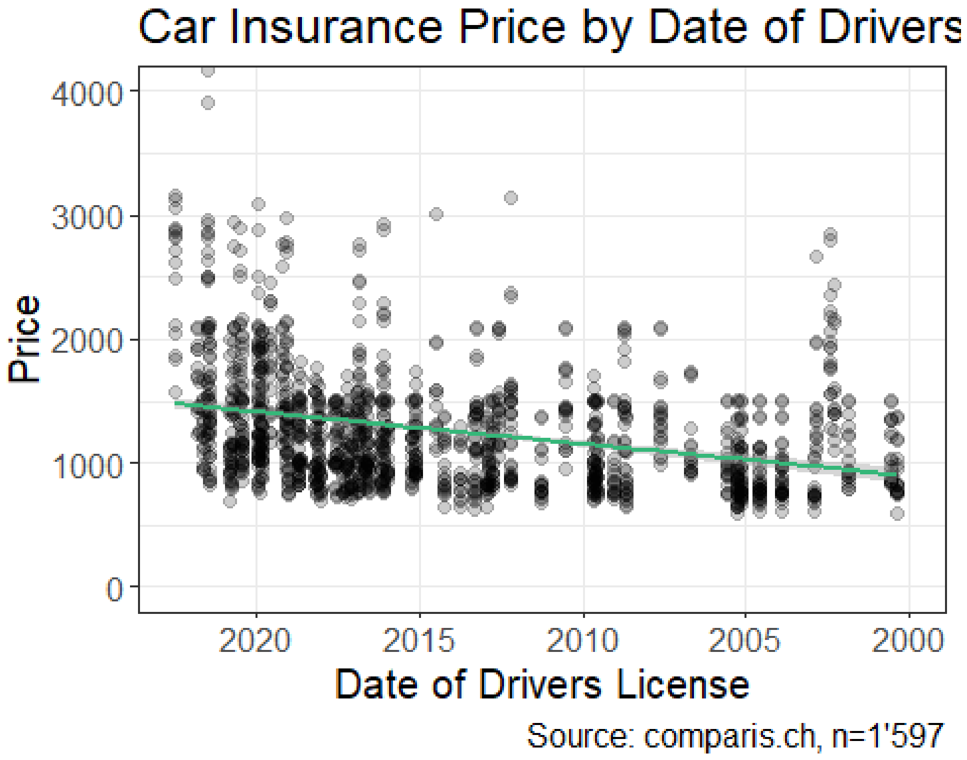
Age and date of drivers license

```
age_quad_plot <- ggplot(data_insurance, aes(x=age, y=price))+
        geom_jitter(size= 1.5, alpha = 0.3, width = 0.5, height = 0)+
        stat_smooth(aes(y=price), method = 'lm', formula = y ~ x + I(x^2), si
ze = 1, color = viridis(4)[4])+
  theme_bw()+
  ylim(0, 4000)+
  labs(x="Age", y="Price",
       title = "Car Insurance Price by Age",
       caption = "Source: comparis.ch, n=2'007")+
  theme(text = element_text(size = 15))

age_quad_plot
```



Source: comparis.ch, n=2'007

```
ggsave("Export/desc_age.png", plot = age_quad_plot, width = 8, height = 6)


# filter date range, since effect is clearer visible and less distortion from
age

tsdl_plot <- data_insurance %>%
  filter(date_of_drivers_license > as.Date('2000-01-01')) %>%
  ggplot(aes(x=date_of_drivers_license, y=price))+
  geom_jitter(size=2, alpha = 0.2, height = 0)+
  theme_bw()+
  geom_smooth(method = "lm", color = viridis(4)[3])+
  coord_x_date(xlim = c(max(data_insurance$date_of_drivers_license), min(as.D
ate('2000-01-01'))), ylim = c(0, 4000))+
  labs(x="Date of Drivers License", y="Price",
       title = "Car Insurance Price by Date of Drivers License",
       caption = "Source: comparis.ch, n=1'597")+
  theme(text = element_text(size = 15))

tsdl_plot
```



Car Insurance Price by Date of Drivers

Source: comparis.ch, n=1'597

```
ggsave("Export/desc_tsdl.png", plot = tsdl_plot, width = 8, height = 6)

#possible logarithmic funciton?
```

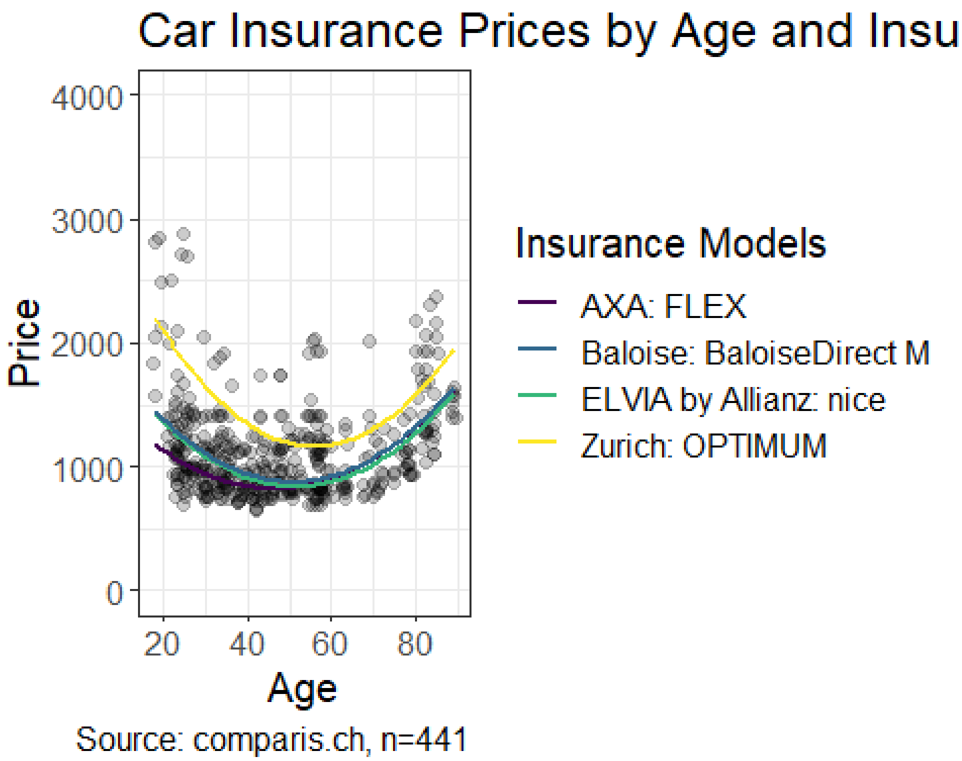
Regression model

```
linear_reg <- lm(data = data_insurance, formula =  price ~ nationality + age
+ gender + time_since_dl)
#summary(linear_reg)

data_insurance$age2 <- data_insurance$age ^ 2

alter_quad_model <- lm(data = data_insurance, price ~ age + age2 + gender + n
ationality + time_since_dl)
#stargazer(alter_quad_model, type = "latex", title = "Regression Analysis", o
ut = "Export/reg_table.latex")
```

Comparing Insurance Models

```
# get most popular models and create data-set
model_count <- data_insurance %>%
  count(insurance_model, sort = T)
model_count

## # A tibble: 19 x 2
##    insurance_model             n
##    <chr>                   <int>
##  1 FLEX                      111
##  2 FLEX Plus                 111
##  3 nice                      111
##  4 PREMIUM                   111
##  5 OPTIMUM                   110
##  6 BaloiseDirect L           109
##  7 BaloiseDirect M           109
##  8 Custom                    109
##  9 fair                      109
## 10 L - Von Allem das Beste   109
## 11 M - Die smarte Wahl       109
## 12 Günstige Basisdeckung     104
## 13 Individueller Schutz      104
## 14 Car L                     102
## 15 Car M                     102
## 16 Comfort                   102
## 17 clever                     96
## 18 premium                    96
## 19 Autoversicherung PLUS      93
```

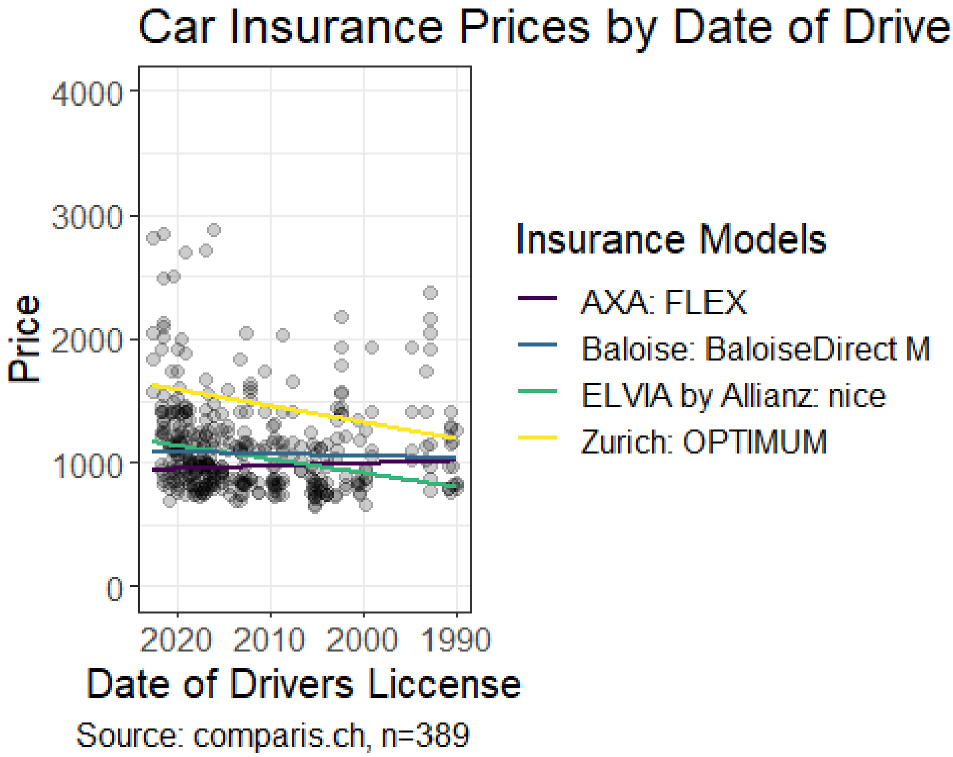
```
select_models <- data_insurance %>%
  filter(insurance_model == "FLEX" |
         insurance_model == "nice" |
         insurance_model == "OPTIMUM" |
         insurance_model == "BaloiseDirect M")

select_models$model_company <- paste0(select_models$insurance_company, ": ",
select_models$insurance_model)

cols <- viridis(4)

# Price by Age by insurance model

age_comparison_plot <- ggplot(select_models, aes(x=age, y=price))+
  geom_jitter(alpha=0.2, size=2, width = 0.5, height = 0)+
  stat_smooth(data = filter(select_models, insurance_model == "FLEX"),
              aes(x=age, y=price, colour = "AXA: FLEX"), method = 'lm', formu
la = y ~ x + I(x^2), size = 1, se = F)+
  stat_smooth(data = filter(select_models, insurance_model == "nice"),
              aes(x=age, y=price, colour = "ELVIA by Allianz: nice"), method
= 'lm', formula = y ~ x + I(x^2), size = 1, se = F)+
  stat_smooth(data = filter(select_models, insurance_model == "OPTIMUM"),
              aes(x=age, y=price, colour = "Zurich: OPTIMUM"), method = 'lm',
formula = y ~ x + I(x^2), size = 1, se = F)+
  stat_smooth(data = filter(select_models, insurance_model == "BaloiseDirect
M"),
              aes(x=age, y=price, colour = "Baloise: BaloiseDirect M"), metho
d = 'lm', formula = y ~ x + I(x^2), size = 1, se = F)+
  scale_color_manual(name="Insurance Models", values = cols)+
  theme_bw()+
  coord_cartesian(ylim = c(0, 4000))+
  labs(x="Age", y="Price",
       title = "Car Insurance Prices by Age and Insurance Model",
       caption = "Source: comparis.ch, n=441")+
  theme(text = element_text(size = 15))
age_comparison_plot
```



Car Insurance Prices by Age and Insu

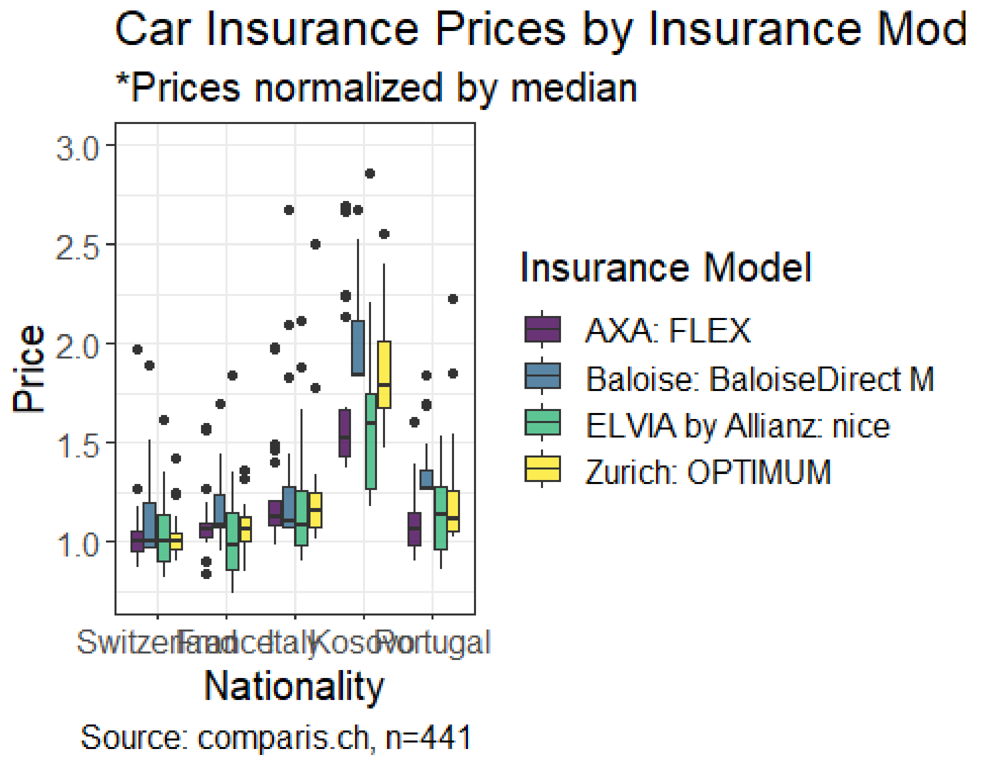
Source: comparis.ch, n=441

```
ggsave("Export/comp_age.png", plot = age_comparison_plot, width = 8, height =
6)


# Price by date of drivers license by insurance model

select_models_filter <- filter(select_models, date_of_drivers_license > as.Da
te('1990-01-01'))

tsdl_comparison_plot <- select_models_filter %>%
  ggplot(aes(x=date_of_drivers_license, y=price))+
    geom_jitter(alpha=0.2, size=2, width = 0.5, height = 0)+
    stat_smooth(data = filter(select_models_filter, insurance_model == "FLEX"
),
                aes(x=date_of_drivers_license, y=price, colour = "AXA: FLEX")
, method = 'lm', size = 1, se = F)+
    stat_smooth(data = filter(select_models_filter, insurance_model == "nice"
),
                aes(x=date_of_drivers_license, y=price, colour = "ELVIA by Al
lianz: nice"), method = 'lm', size = 1, se = F)+
    stat_smooth(data = filter(select_models_filter, insurance_model == "OPTIM
UM"),
                aes(x=date_of_drivers_license, y=price, colour = "Zurich: OPT
IMUM"), method = 'lm', size = 1, se = F)+
    stat_smooth(data = filter(select_models_filter, insurance_model == "Baloi
seDirect M"),
                aes(x=date_of_drivers_license, y=price, colour = "Baloise: Ba
loiseDirect M"), method = 'lm', size = 1, se = F)+
    scale_color_manual(name="Insurance Models", values = cols)+
    theme_bw()+
    coord_x_date(xlim = c(max(select_models_filter$date_of_drivers_license),
                          min(select_models_filter$date_of_drivers_license)),
ylim = c(0, 4000))+
    labs(x="Date of Drivers Liccense", y="Price",
         title = "Car Insurance Prices by Date of Drivers License and Insuran
ce Model",
         caption = "Source: comparis.ch, n=389")+
  theme(text = element_text(size = 15))
tsdl_comparison_plot
```



```
ggsave("Export/comp_tsdl.png", plot = tsdl_comparison_plot, width = 8, height
= 6)

### Price by nationality by insurance model

# calculate medians for nationality Switzerland for each model
median_model_swiss <- select_models %>% filter(nationality == "Switzerland")
%>%
  group_by(insurance_model) %>%
  summarise(median=median(price))

# get normalized prices by deviding through median
```

```
select_models <- select_models %>%
  mutate(price_corr = as.numeric(ifelse(insurance_model=="BaloiseDirect M", p
rice/median_model_swiss$median[1],
                             ifelse(insurance_model=="FLEX", price/median_mod
el_swiss$median[2],
                                    ifelse(insurance_model=="nice", price/med
ian_model_swiss$median[3],
                                           ifelse(insurance_model=="OPTIMUM",
price/median_model_swiss$median[4], "no"))))))

corrected_plot2 <- ggplot(select_models, aes(x=nationality, y=price_corr, fil
l=model_company))+
    geom_boxplot()+
    theme_bw()+
    labs(y="Price", x="Nationality", fill = "Insurance Model",
         title = "Car Insurance Prices by Insurance Models and Nationality",
         subtitle = "*Prices normalized by median",
         caption = "Source: comparis.ch, n=441")+
  coord_cartesian(ylim = c(0.75, 3))+
  scale_fill_viridis(discrete = T, alpha = 0.8)+
  theme(text = element_text(size = 15))
corrected_plot2
```


Car Insurance Prices by Insurance Mod

*Prices normalized by median

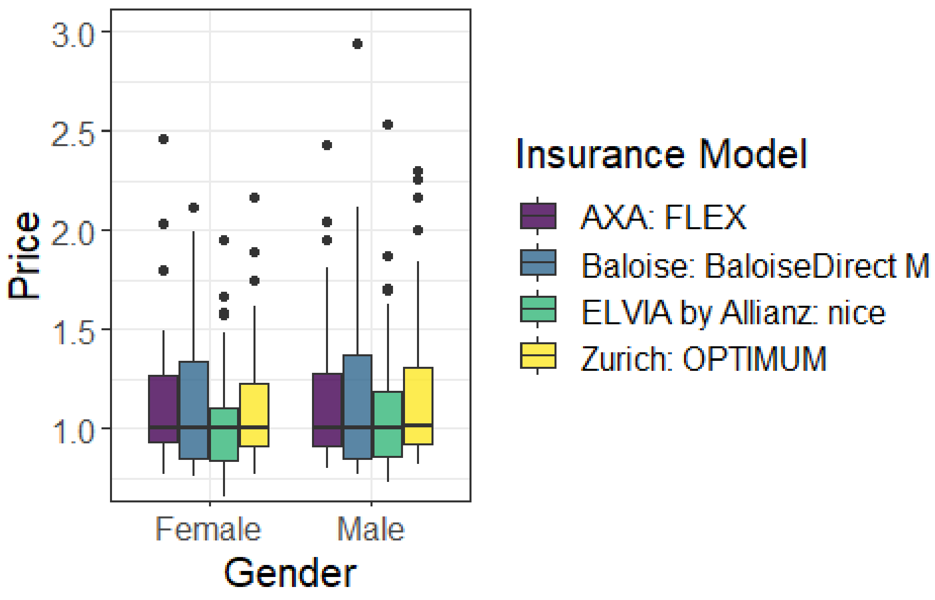
Source: comparis.ch, n=441

```
ggsave("Export/comp_nationality.png", plot = corrected_plot2, width = 8, heig
ht = 6)
```

```
### Price by gender by insurance model

# calculate medians for gender Female for each model

median_model_gender <- select_models %>% filter(gender == "Female") %>%
  group_by(insurance_model) %>%
  summarise(median=median(price))

# get normalized prices by deviding through median
select_models <- select_models %>%
  mutate(price_corr_gender = as.numeric(ifelse(insurance_model=="BaloiseDirec
t M", price/median_model_gender$median[1],
                             ifelse(insurance_model=="FLEX", price/median_mod
el_gender$median[2],
                                    ifelse(insurance_model=="nice", price/med
ian_model_gender$median[3],
                                           ifelse(insurance_model=="OPTIMUM",
price/median_model_gender$median[4], "no"))))))

corrected_plot_gender <- ggplot(select_models, aes(x=gender, y=price_corr_gen
der, fill=model_company))+
    geom_boxplot()+
    theme_bw()+
    labs(y="Price", x="Gender", fill = "Insurance Model",
         title = "Car Insurance Prices by Insurance Models and Gender",
         subtitle = "*Prices normalized by median",
         caption = "Source: comparis.ch, n=441")+
  coord_cartesian(ylim = c(0.75, 3))+
  scale_fill_viridis(discrete = T, alpha = 0.8)+
  theme(text = element_text(size = 15))
corrected_plot_gender
```

Car Insurance Prices by Insurance Mod

*Prices normalized by median



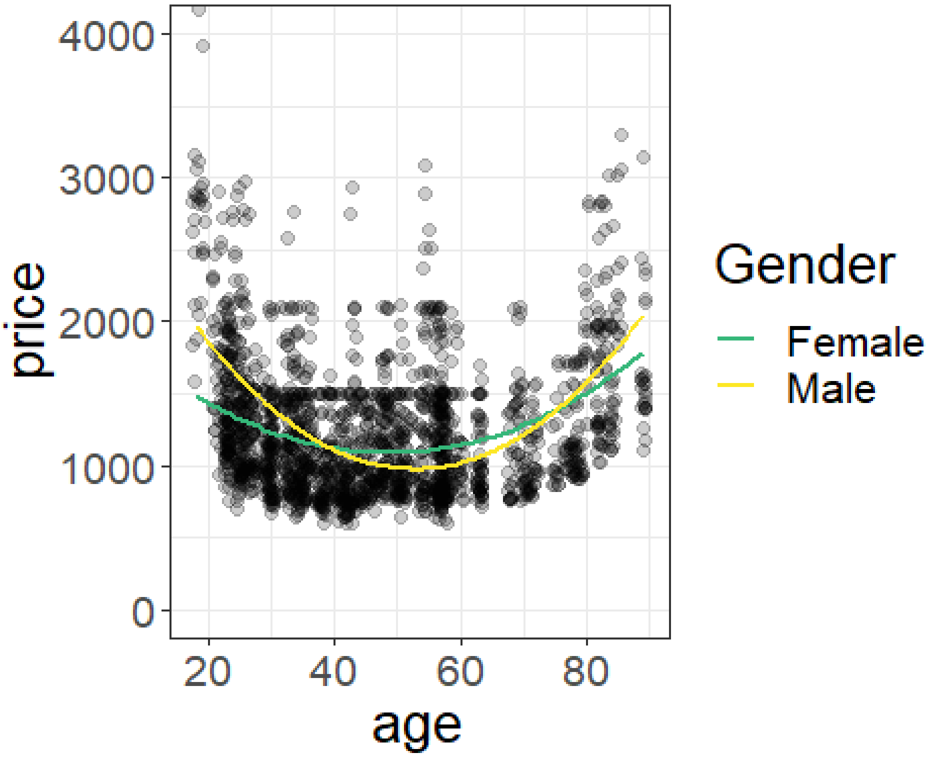
Source: comparis.ch, n=441

```
ggsave("Export/comp_gender.png", plot = corrected_plot_gender, width = 8, hei
ght = 6)
```

Possible further analysis

Prices by age and gender

```
cols <- viridis(4)[3:4]

age_gender_plot <- ggplot(data_insurance, aes(x=age, y=price))+
  geom_jitter(alpha=0.2, size=2, width = 0.5, height = 0)+
  stat_smooth(data = filter(data_insurance, gender == "Female"),
              aes(x=age, y=price, colour = "Female"), method = 'lm', formula
= y ~ x + I(x^2), size = 1, se = F, fullrange = T)+
  stat_smooth(data = filter(data_insurance, gender == "Male"),
              aes(x=age, y=price, colour = "Male"), method = 'lm', formula =
y ~ x + I(x^2), size = 1, se = F, fullrange = T)+
  scale_color_manual(name="Gender", values = cols)+
  theme_bw()+
  coord_cartesian(ylim = c(0, 4000))+
  theme(text = element_text(size = 20))

age_gender_plot
```


4000
3000
2000
1000
0
price
20
40
60
80
age
Gender
Female
Male