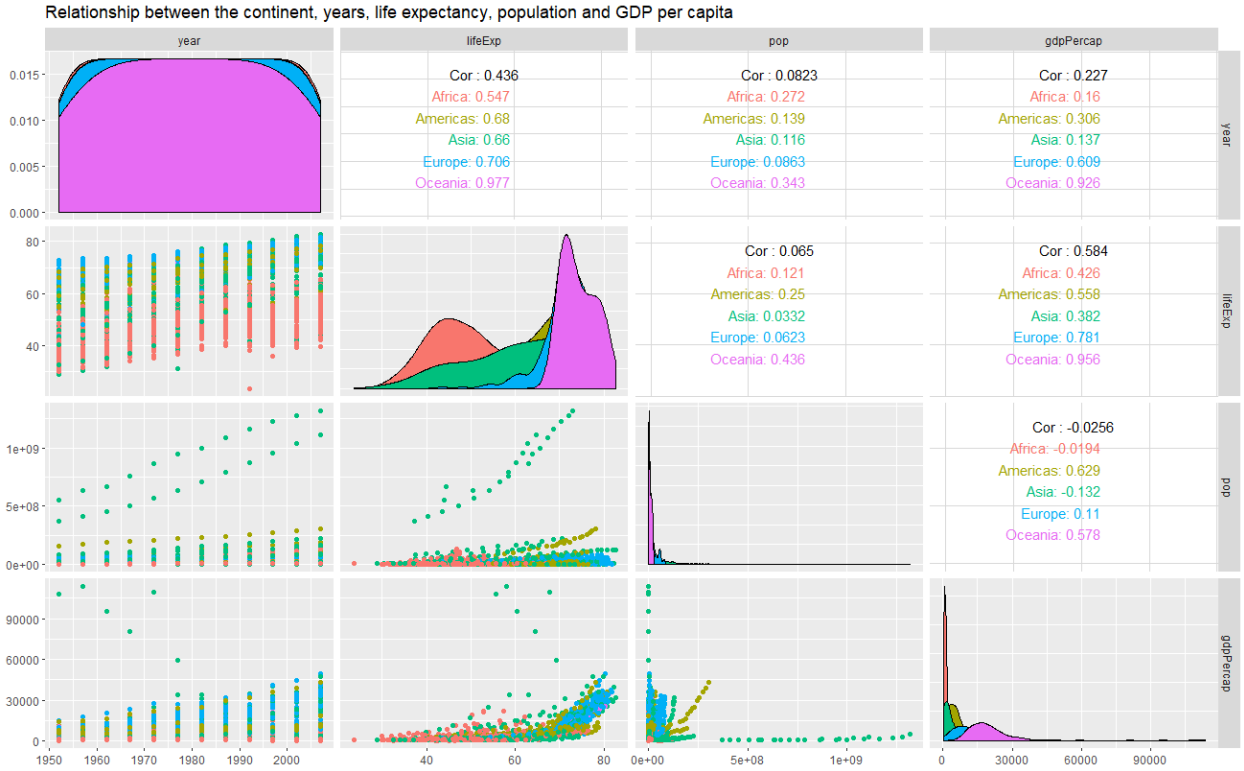
STA 3000 - Statistical computing
In-class activity 2

Gapminder

Create a figure that shows the relationship between the continent, year, life expectancy, population, and GDP per capita. Your figure can contain more than one plot / facet / panel. Interpret in detail the relationships that you see in the plots. Make sure that the labels and the title are interpretable.

```
gm = gapminder %>% select(year, lifeExp, pop, gdpPercap)
```

```
ggpairs(gm, title = "Relationship between the continent, years, life expectancy, population and GDP per capita", ggplot2::aes(colour=gapminder$continent))
```



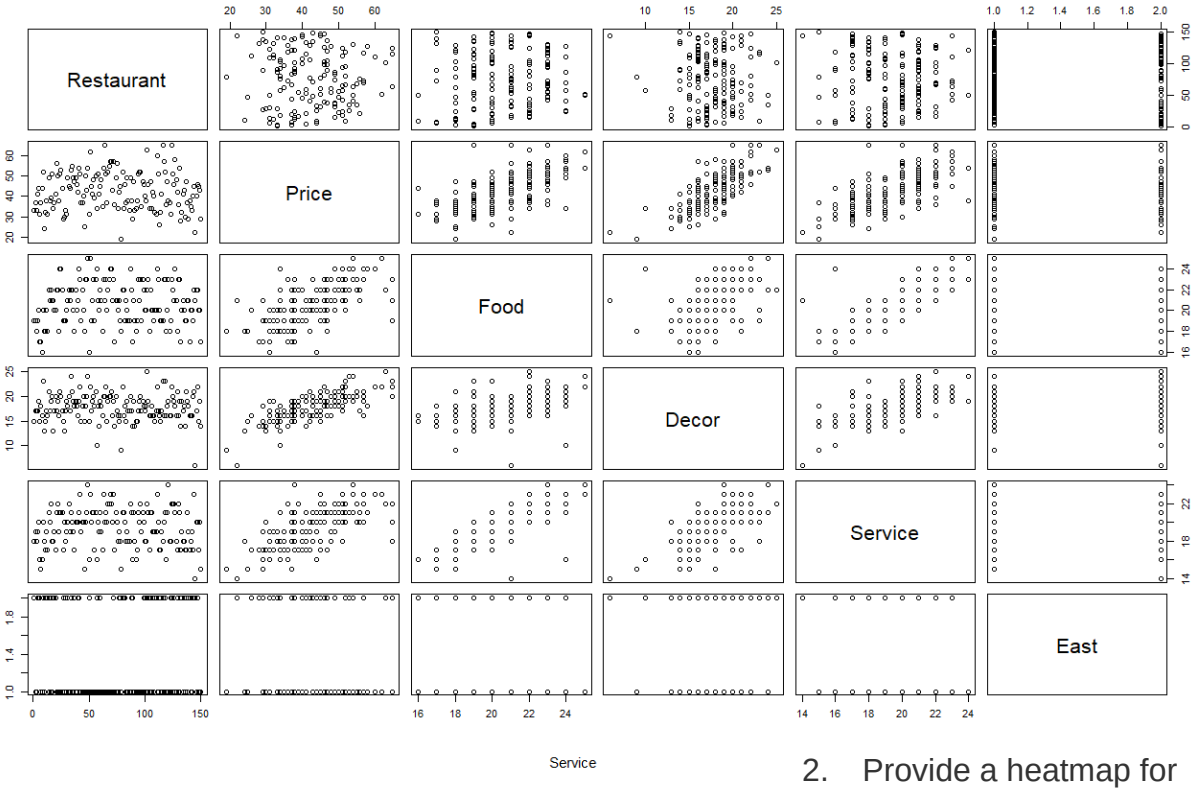
We can see from the ggpairs plot below that: our life expectation is gradually going up these years, Europe and Americas are top 2 continents with the highest life expectation compared to other continents, Africa is the worst; As for population changes over years, the overall population is increasing mostly because of the growth of Asia's population; From 1950 to 2009, GDP per capita overall is growing overtime, Europe and America's GDP per capita are relatively higher, one country in Asia had a way higher capita GDP

before 1980; Life expectation's relationships with population and GDP per capita are positive; There's no clear relationship between population and GDP per capita.

Italian Restaurants in NYC

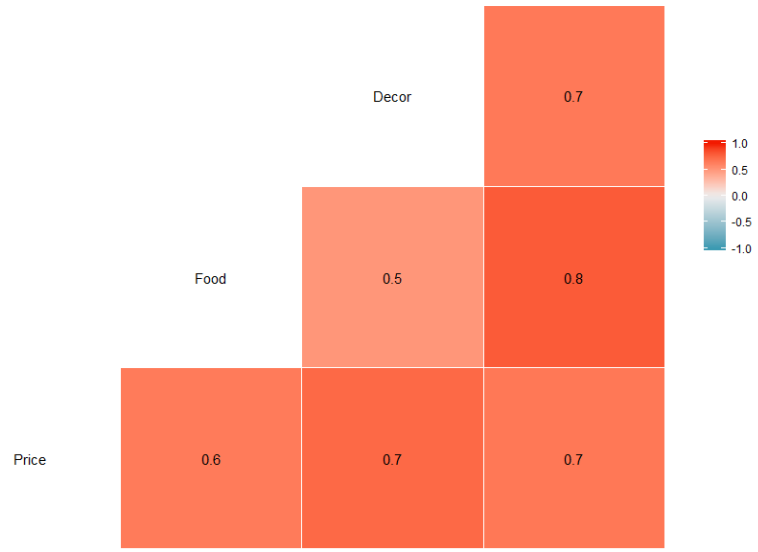
1. Create a figure that contains plots for all the pairs of variables in the dataset, except Case (i.e., a figure that contains plots for Restaurant vs Price, Food vs Price, Decor vs Service, etc.). Describe what you see in the plots. What are the strongest and weakest relationships you see?

```
pairs(itlRest[2:7], cex.labels=2)
```



2. Provide a heatmap for the correlation between the numerical variables in the dataset. What can you see?

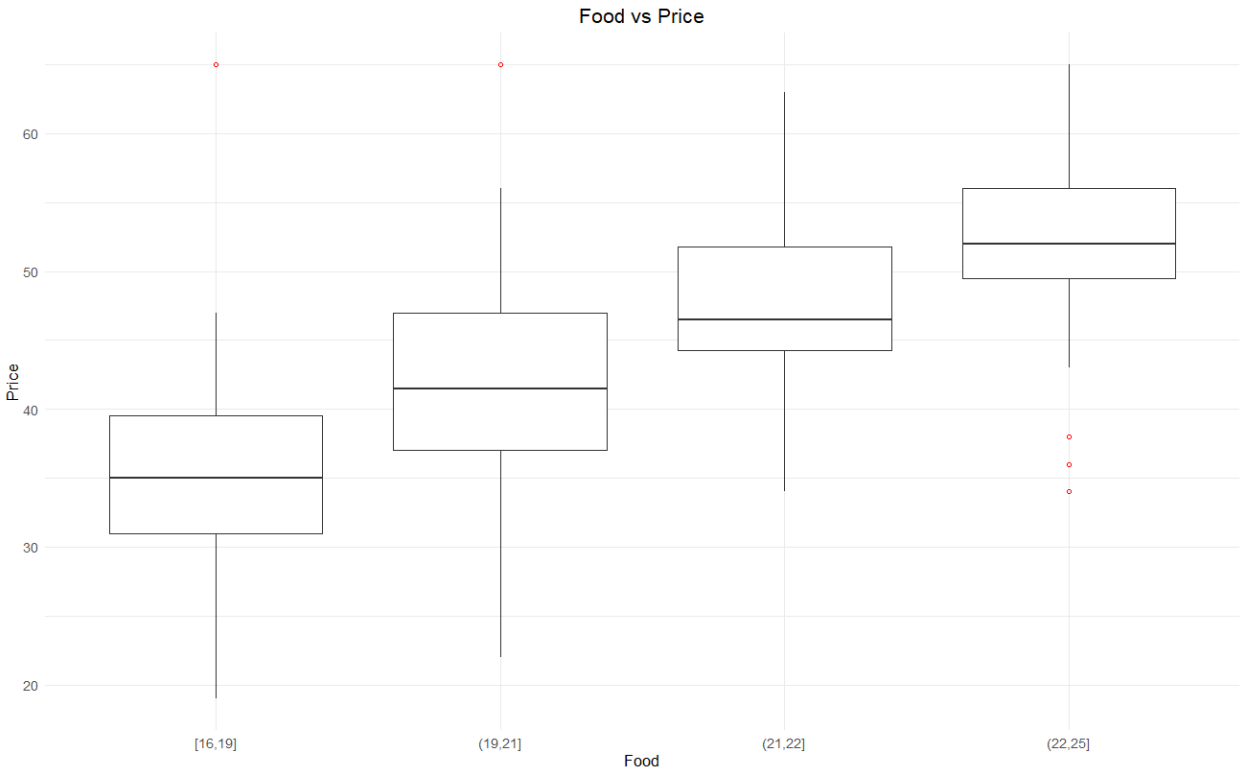
```
numeric = itlRest %>%
select (3,4,5,6)
ggcorr(numeric, label =
TRUE)
```



We can see from the heatmap that the strongest linear correlation is between service and food; the lowest linear correlation is between décor and food; all these correlations are positive.

3. Find 2 examples of cheap restaurants that have relatively good food and 2 examples of expensive restaurants that have relatively bad food.

```
itlRest$Food = cut(itlRest$Food,
                        breaks=quantile(itlRest$Food),
                        include.lowest = TRUE)

ggplot(itlRest) +
  aes(x = Food, y = Price) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 1) +
  ggtitle("Food vs Price") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(text=element_text(size=13))

itlRest %>%
  select(Food, Price, Restaurant) %>%
  arrange(Food, Price)
```



I quantify food quality and plot a boxplot between the food quality and price. It's clear to see that those outliers below the "box" whose food quality are lower than 21 are examples with relatively higher price and relatively bad food; those outliers below the "box" whose food quality are above 21 are examples with relatively lower price and relatively good food.

	Food	Price	Restaurant
1	[16,19]	19	Lamarca
2	[16,19]	24	Bardolino
3	[16,19]	25	Ecco-la
4	[16,19]	28	Supreme Macaroni Co.
5	[16,19]	29	Mangia e Bevi
6	[16,19]	29	Cara Mia
7	[16,19]	29	Zucchero e Pomodori
8	[16,19]	30	Carino Ristorante
9	[16,19]	30	Tony's Di Napoli
10	[16,19]	31	Baci
11	[16,19]	31	Bella Luna
12	[16,19]	31	Ernie's
13	[16,19]	31	Sambuca, Trattoria
14	[16,19]	32	Puttanesca
15	[16,19]	32	Basta Pasta
16	[16,19]	33	Maruzzella
17	[16,19]	33	Amarone
18	[16,19]	33	Luna Piena
19	[16,19]	33	Casa Di Meglio
20	[16,19]	33	Andiamo
21	[16,19]	34	Mediterraneo
22	[16,19]	34	Pappardella
23	[16,19]	35	Grace's Trattoria
24	[16,19]	35	Notaro
25	[16,19]	36	Ribollita
26	[16,19]	36	Il Vagabondo
27	[16,19]	37	Anche Vivolo
28	[16,19]	37	Divino Ristorante
29	[16,19]	37	Sal Anthony's
30	[16,19]	37	Baraonda
31	[16,19]	38	Mezzaluna
32	[16,19]	38	Caffe Grazie
33	[16,19]	38	Trattoria Dopo Teatro
34	[16,19]	38	Luna Blu
35	[16,19]	39	Fred's at Barneys NY
36	[16,19]	40	Paul & Jimmy's
37	[16,19]	40	Tuscan Square
38	[16,19]	41	Arno
39	[16,19]	41	Quattro Gatti
40	[16,19]	42	Sette MoMA
41	[16,19]	44	Barbaresco
42	[16,19]	45	La Rivista
43	[16,19]	45	Fino
44	[16,19]	46	Limoncello
45	[16,19]	47	Torre di Pisa
46	[16,19]	47	Coco Pazzo Teatro
47	[16,19]	65	Rainbow Grill
48	(19,21]	22	Veronica
49	(19,21]	26	Puccini
50	(19,21]	29	Trattoria Del Sogno
51	(19,21]	31	Pomodoro Rosso
52	(19,21]	31	Casa Mia

53	(19,21]	32	Tello's Ristorante
54	(19,21]	32	Pietrasanta
55	(19,21]	33	Triangolo
56	(19,21]	34	Pasticcio
57	(19,21]	34	La Gioconda
58	(19,21]	34	Biricchino
59	(19,21]	35	Trattoria Rustica
60	(19,21]	37	Trattoria Alba
61	(19,21]	37	BondÃ RistoranteÂ´
62	(19,21]	37	Il Gatto & La Volpe
63	(19,21]	37	Luca
64	(19,21]	38	Becco
65	(19,21]	38	Bello
66	(19,21]	38	Osteria Laguna
67	(19,21]	40	Osteria al Doge
68	(19,21]	40	East River Cafe
69	(19,21]	41	Bottino
70	(19,21]	41	Firenze
71	(19,21]	41	Teodora
72	(19,21]	41	Gino
73	(19,21]	42	Da Tommaso
74	(19,21]	42	Patsy's
75	(19,21]	42	La Grolla, Ristorante
76	(19,21]	43	Il Riccio
77	(19,21]	43	Vivolo
78	(19,21]	43	Tino's
79	(19,21]	44	Cinque Terre
80	(19,21]	44	Artusi
81	(19,21]	45	Vico
82	(19,21]	46	Scaletta
83	(19,21]	46	Da Filippo
84	(19,21]	47	Rossini's
85	(19,21]	47	Marchi's
86	(19,21]	48	Giovanni Venticinque
87	(19,21]	48	Sette Mezzo
88	(19,21]	49	Lumi
89	(19,21]	49	Castellano
90	(19,21]	49	Coco Pazzo CafÃ©
91	(19,21]	49	Il Valentino
92	(19,21]	51	Bice
93	(19,21]	51	Giambelli
94	(19,21]	52	Barbetta
95	(19,21]	52	Nicola Paone
96	(19,21]	56	Il Valletto Due Mila
97	(19,21]	65	Harry Cipriani
98	(21,22]	34	Le Zie
99	(21,22]	37	Due
100	(21,22]	38	MÃ©tisse
101	(21,22]	39	Belluno
102	(21,22]	40	Via Oreto
103	(21,22]	44	Canaletto
104	(21,22]	45	Vago Ristorante
105	(21,22]	45	Giovanni
106	(21,22]	45	Cellini
107	(21,22]	46	Orso
108	(21,22]	46	ViceVersa
109	(21,22]	47	Parma
110	(21,22]	49	Da Antonio Ristorante
111	(21,22]	49	Bellini
112	(21,22]	50	Bruno Ristorante
113	(21,22]	51	San Giusto
114	(21,22]	52	Le Madri
115	(21,22]	52	Il Menestrello
116	(21,22]	53	Circo Osteria del

```
117 (21,22]    55            Coco Pazzo
118 (21,22]    56          Bravo Gianni
119 (21,22]    63                 Palio
120 (22,25]    34               Gennaro
121 (22,25]    36          Sirabella's
122 (22,25]    38              Rughetta
123 (22,25]    43     Enoteca i Trulli
124 (22,25]    46              Paola's
125 (22,25]    47              NovitÃ¡
126 (22,25]    48            Follonico
127 (22,25]    49            Lusardi's
128 (22,25]    50     Fresco by Scotto
129 (22,25]    50                Elio's
130 (22,25]    51             Campagna
131 (22,25]    51             I Trulli
132 (22,25]    51              Nanni's
133 (22,25]    51              Primola
134 (22,25]    51             DeGrezia
135 (22,25]    52                Nino's
136 (22,25]    52                  Remi
137 (22,25]    53           Campagnola
138 (22,25]    54           Da Umberto
139 (22,25]    54              Sistina
140 (22,25]    54              Grifone
141 (22,25]    54              Erminia
142 (22,25]    55           Il Monello
143 (22,25]    57                Rao's
144 (22,25]    57           Il Postino
145 (22,25]    57           Il Tinello
146 (22,25]    57              Il Nido
147 (22,25]    58         Scalinatella
148 (22,25]    60            Primavera
149 (22,25]    62              FELIDIA
150 (22,25]    65         San Domenico
```

As I highlighted below, the two restaurants with relatively higher price and relatively bad food are Rainbow Grill and Harry Cipriani; the two restaurants with relatively lower price and relatively good food are Gennaro and Sirabella's.

Although here I only list the examples that can outliers, technically any examples with the below-average food quality and above-average price could be considered as examples with relatively higher price and relatively bad food; vice versa for examples with relatively lower price and relatively good food.

4. Suppose you're going on a date and want to use the information in this dataset to pick where to go. Assume your budget is at most $40. Assuming that you can get a table anywhere you want, where would you go and why?

   First I created a new column as the mean value of the standardized value of price, food, décor and service four variables.

```
Price <- (itlRest$Price - mean(itlRest$Price))/sd(itlRest$Price)
Food <- (itlRest$Food - mean(itlRest$Food))/sd(itlRest$Food)
Decor <- (itlRest$Decor - mean(itlRest$Decor))/sd(itlRest$Decor)
```

```
Service <- (itlRest$Service -
mean(itlRest$Service))/sd(itlRest$Service)
itlRest$rating = (Price+Food+Decor+Service)/4

itlRest %>%
  filter(Price <= 40) %>%
  arrange(desc(rating)) %>%
  select(Restaurant, rating) %>%
  top_n(1)

  Restaurant    rating
1   Rughetta 0.8453116
```

I would go to restaurant Rughetta under the assumption that price, food, decoration, service these four factors have the same weight.

5. Create a figure that displays the relationship between price, food, decor, service, and the East / West indicator. Your figure can contain more than one plot / facet / panel. Make sure that the labels and the title are interpretable. Interpret in detail the relationships that you see.

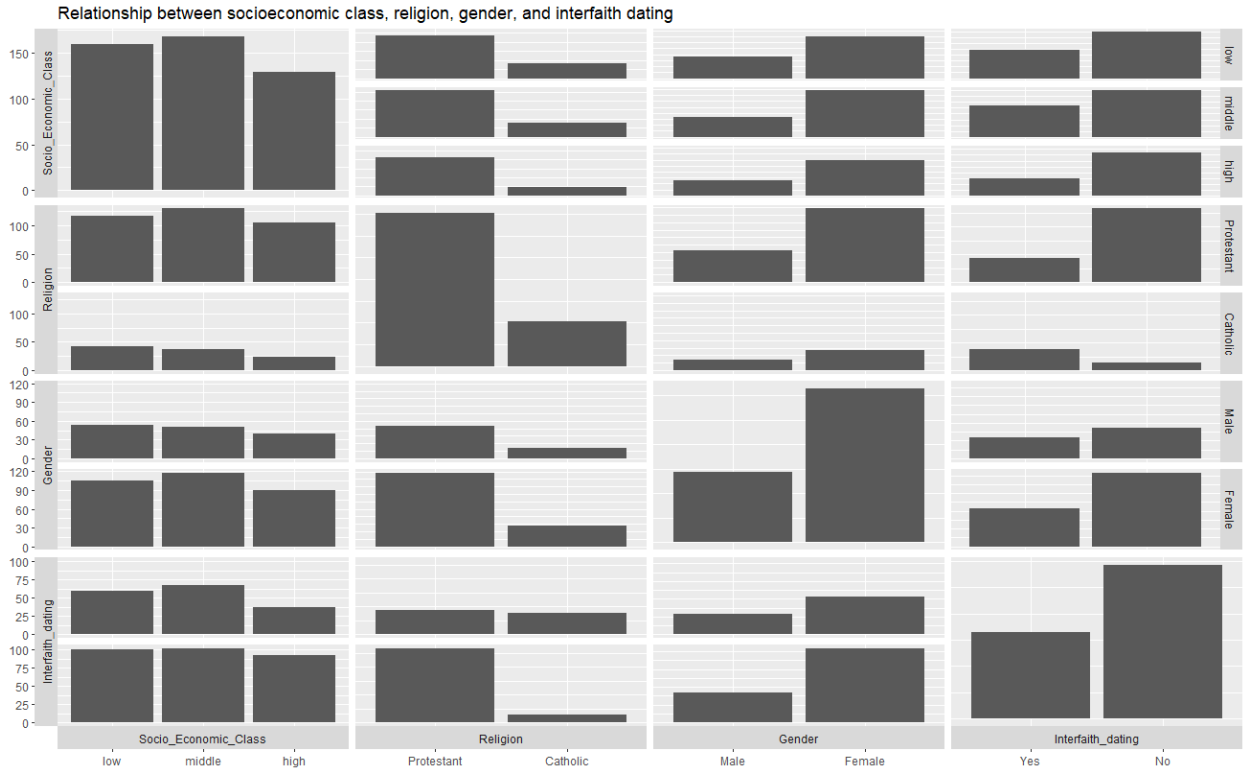
```
ggpairs(itlRest, columns = c(3:6), title = "Relationships between
price, food, decor, service, and the East / West indicator",
ggplot2::aes(colour=itlRest$East))
```

As we could see that the relationships between price, food, décor and service are all positive, the correlations between food and price, food and décor, food and service are higher for east side restaurants.

Interfaith dating data

Create a figure that shows the relationship between socioeconomic class, religion, gender, and the indicator of interfaith dating. Your figure can contain more than one plot / facet / panel. Interpret in detail the relationships that you see in the plots. Make sure that the labels and the title are interpretable.

```
intf =
read.table("http://users.stat.ufl.edu/~winner/data/interfaith.dat")

colnames(intf) <- c("Socio_Economic_Class", "Religion", "Gender",
"Interfaith_dating", "Count")
str(intf)

intf$Socio_Economic_Class = factor(intf$Socio_Economic_Class)
levels(intf$Socio_Economic_Class) =  c("low","middle", "high")

intf$Religion = factor(intf$Religion)
levels(intf$Religion) =  c("Protestant","Catholic")

intf$Gender = factor(intf$Gender)
levels(intf$Gender) =  c("Male","Female")

intf$Interfaith_dating = factor(intf$Interfaith_dating)
levels(intf$Interfaith_dating) =  c("Yes","No")

intf <- intf %>% uncount(Count)

ggpairs(intf, switch = "both", title = "Relationship between
socioeconomic class, religion, gender, and interfaith dating")
```



From the ggpairs plot above, we could see that religion has a strong influence on gender and interfaith dating in that the male Catholics has a relatively higher proportion compared to the Protestants; Catholics have a smaller number of believers than the Protestants; The female tend to join Protestants compared with male; We could also see that Catholics are more open to interfaith fating than the Protestants.