EDA

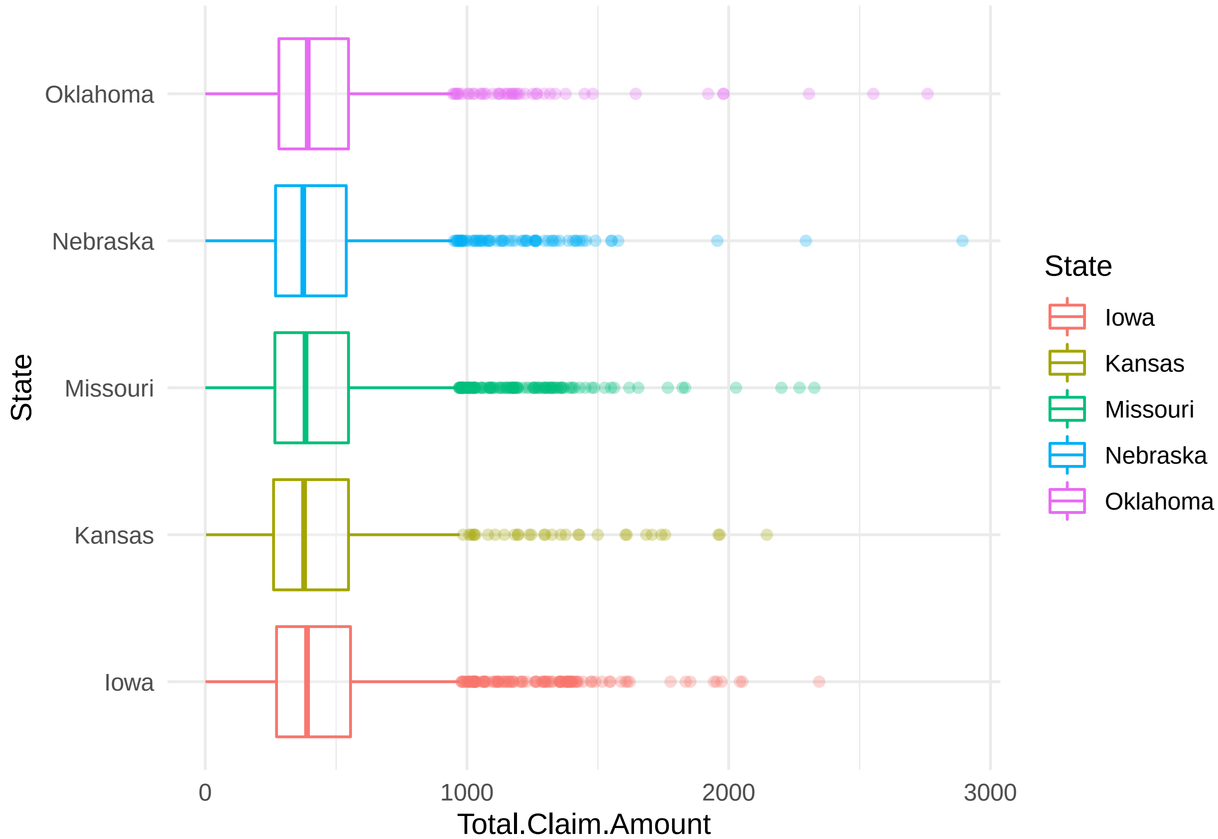
Load data and remove some variables.

```
data = read.csv("~/Downloads/new_train/new_train.csv", stringsAsFactors=TRUE)
# remove State.Code as its the same as State and Country because there is only one
# also remove gender. could lead to biased predictions
data = data %>% select(-c(State.Code, Country, Gender, Customer, Effective.To.Date))
str(data)
```

```
## 'data.frame':    7784 obs. of  20 variables:
##  $ State                        : Factor w/ 5 levels "Iowa","Kansas",..: 2 4 5 3 2 1 1 4 1 1 ...
##  $ Response                     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
##  $ Coverage                     : Factor w/ 3 levels "Basic","Extended",..: 1 2 3 1 1 1 1 3 1 2 ...
##  $ Education                    : Factor w/ 5 levels "Bachelor","College",..: 1 1 1 1 1 1 2 5 1 2 ..
##  $ EmploymentStatus             : Factor w/ 5 levels "Disabled","Employed",..: 2 5 2 5 2 2 2 5 3 2 .
##  $ Income                       : int  56274 0 48767 0 43836 62902 55350 0 14072 28812 ...
##  $ Location.Code                : Factor w/ 3 levels "Rural","Suburban",..: 2 2 2 2 1 1 2 3 2 3 ...
##  $ Marital.Status               : Factor w/ 3 levels "Divorced","Married",..: 2 3 2 2 3 2 2 3 1 2 ..
##  $ Monthly.Premium.Auto         : int  69 94 108 106 73 69 67 101 71 93 ...
##  $ Months.Since.Last.Claim      : int  32 13 18 18 12 14 0 0 13 17 ...
##  $ Months.Since.Policy.Inception: int  5 42 38 65 44 94 13 68 3 7 ...
##  $ Number.of.Open.Complaints    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Number.of.Policies           : int  1 8 2 7 1 2 9 4 2 8 ...
##  $ Policy.Type                  : Factor w/ 3 levels "Corporate Auto",..: 1 2 2 1 2 2 1 1 1 3 ...
##  $ Policy                       : Factor w/ 9 levels "Corporate L1",..: 3 6 6 2 4 6 3 3 3 8 ...
##  $ Claim.Reason                 : Factor w/ 4 levels "Collision","Hail",..: 1 4 1 1 1 2 1 1 1 2 ...
##  $ Sales.Channel                : Factor w/ 4 levels "Agent","Branch",..: 1 1 1 3 1 4 1 1 1 2 ...
##  $ Total.Claim.Amount           : num  385 1131 566 530 138 ...
##  $ Vehicle.Class                : Factor w/ 6 levels "Four-Door Car",..: 6 1 6 5 1 6 1 1 1 1 ...
##  $ Vehicle.Size                 : Factor w/ 3 levels "Large","Medsize",..: 2 2 2 2 2 2 2 2 2 2 ...
```

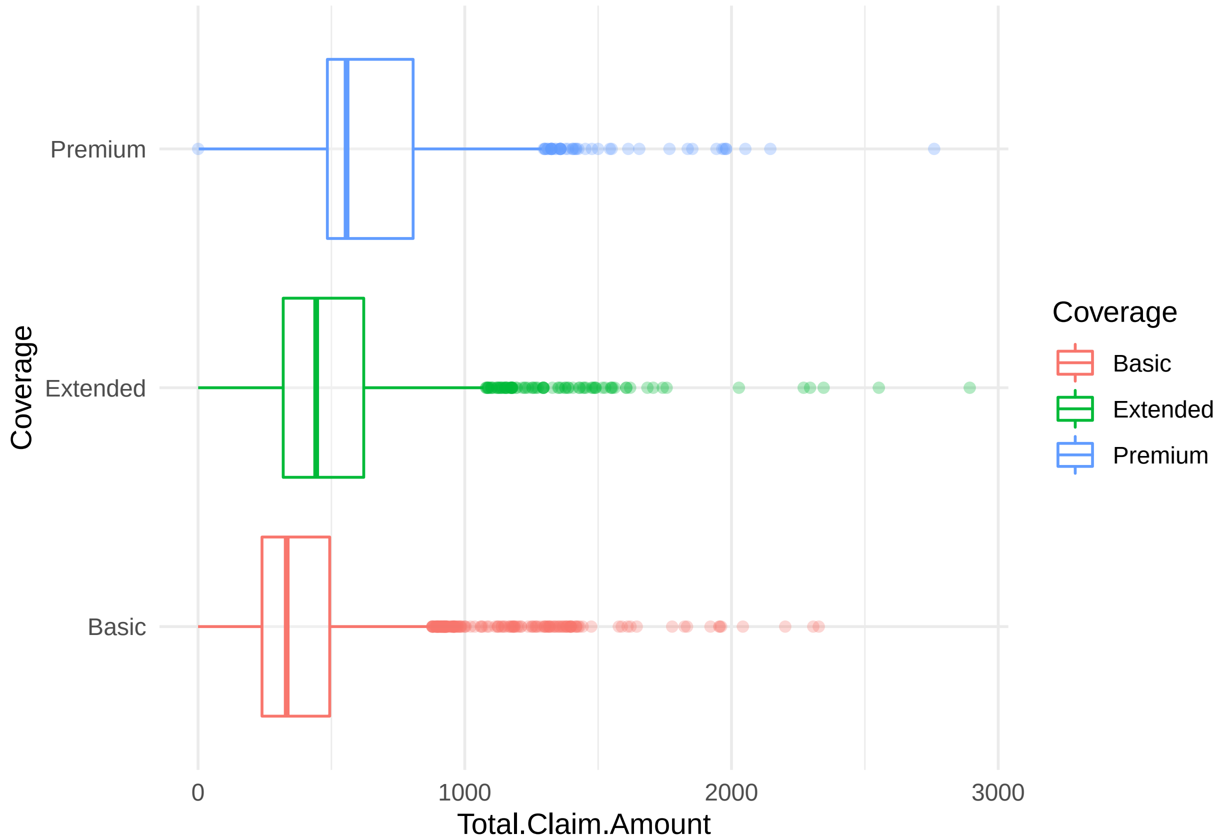
Claims by state. Nothing obvious here.

```
data %>%
  ggplot(aes(y=State,x=Total.Claim.Amount))+
  geom_boxplot(aes(color=State), alpha=0.3)
```



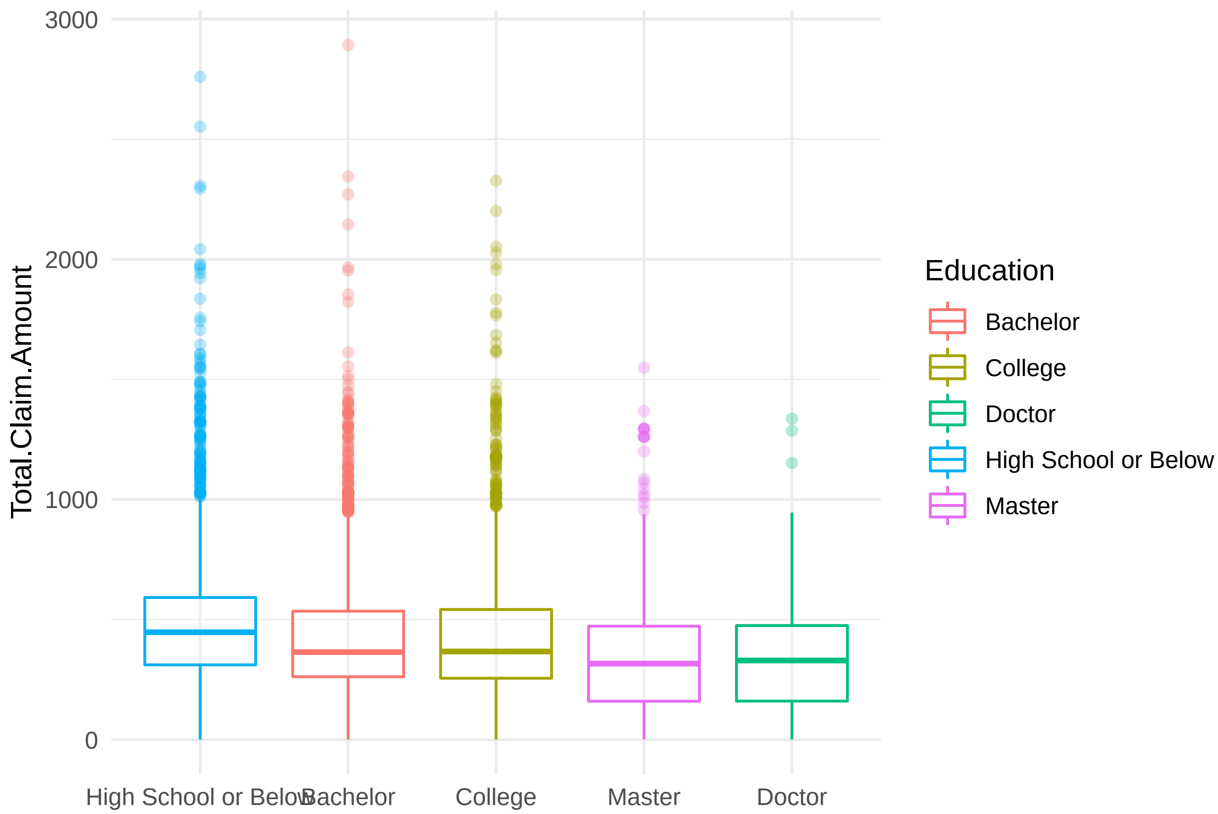
Looking at the coverage types, it is clear that overall having an Extended or Premium leads to larger claims. Makes sense.

```
data %>%
  ggplot(aes(y=Coverage,x=Total.Claim.Amount))+
  geom_boxplot(aes(color=Coverage), alpha=0.3)
```



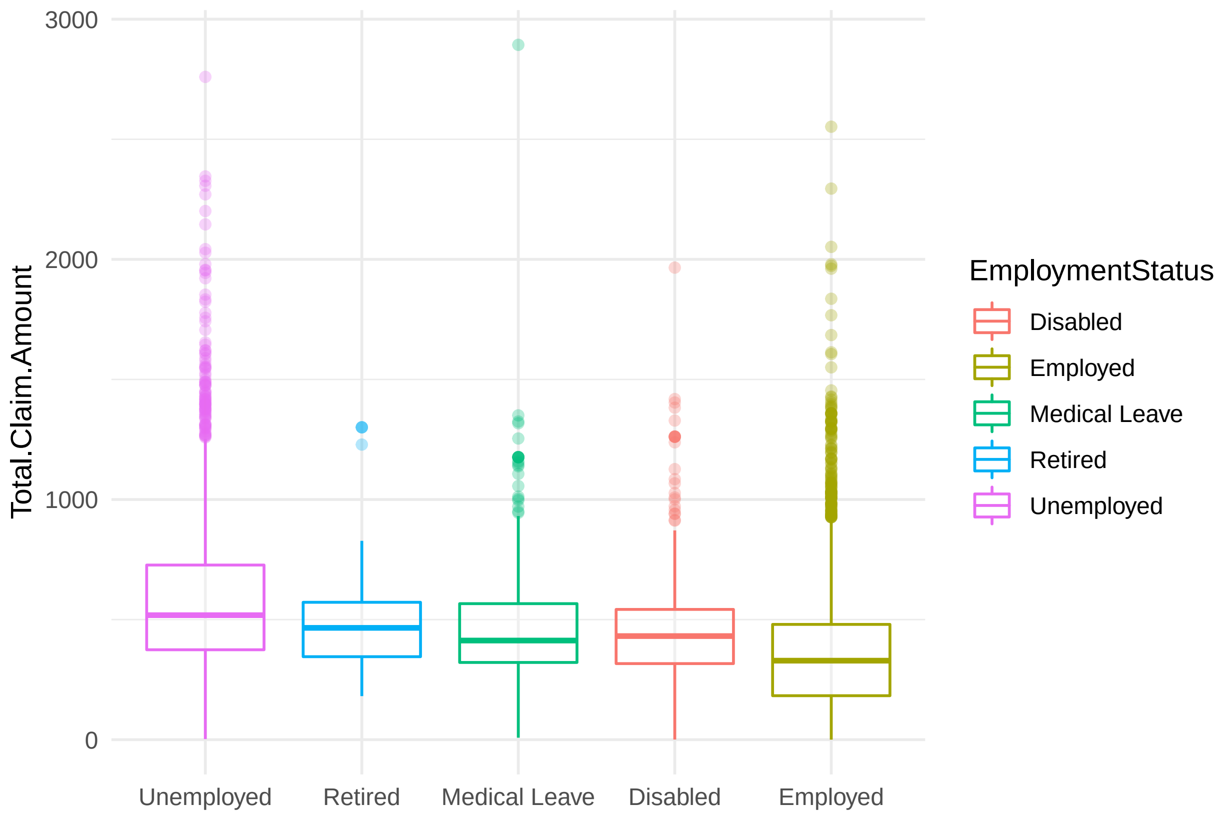
There does seem to be some trend suggesting that the higher the education form one has completed, the lower a claim they will make on average.

```
data %>%
  ggplot(aes(x=reorder(Education, -Total.Claim.Amount),y=Total.Claim.Amount))+
  geom_boxplot(aes(color=Education), alpha=0.3)+
  xlab('')
```



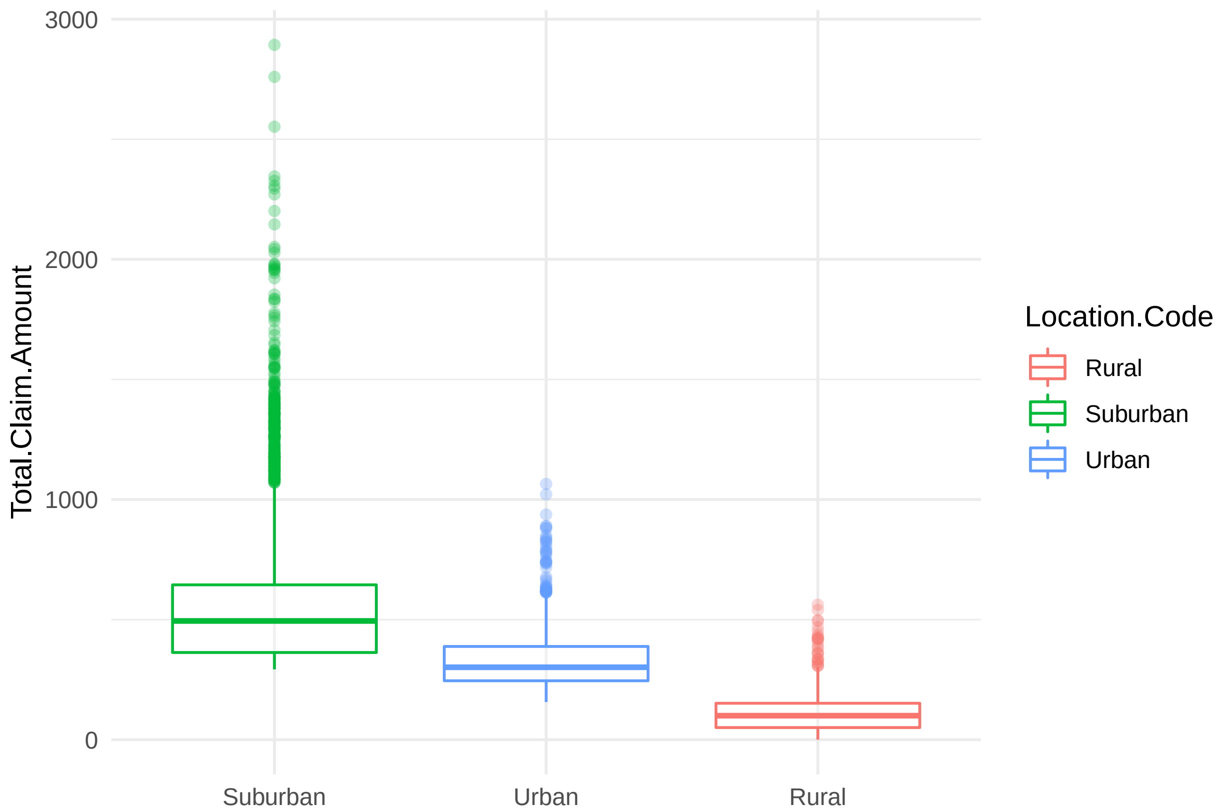
Employment and total claim. Maybe unemployed have slightly higher claims on average.

```
data %>%
  ggplot(aes(x=reorder(EmploymentStatus, -Total.Claim.Amount),y=Total.Claim.Amount))+
  geom_boxplot(aes(color=EmploymentStatus), alpha=0.3)+
  xlab('')
```



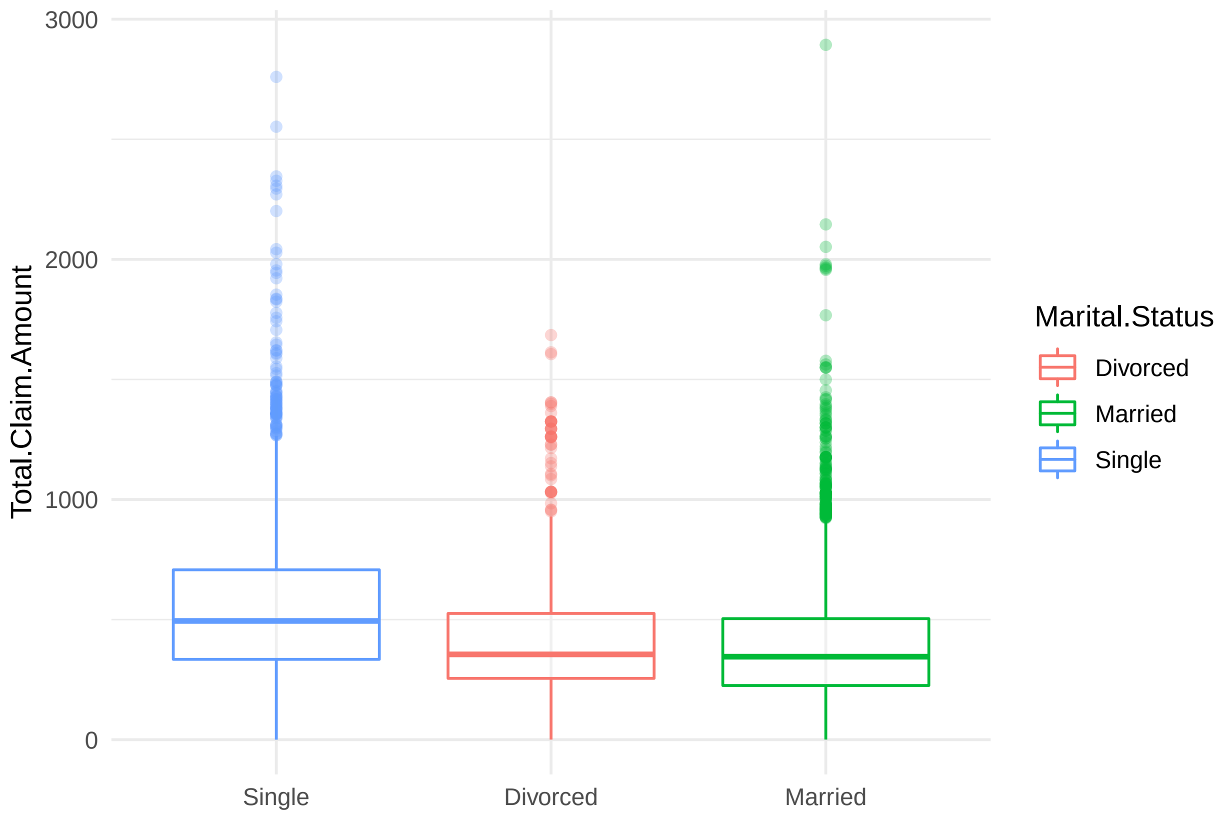
This could again be an important factor. There is quite a big difference between suburban and rural claims.

```
data %>%
  ggplot(aes(x=reorder(Location.Code, -Total.Claim.Amount),y=Total.Claim.Amount))+
  geom_boxplot(aes(color=Location.Code), alpha=0.3)+
  xlab('')
```



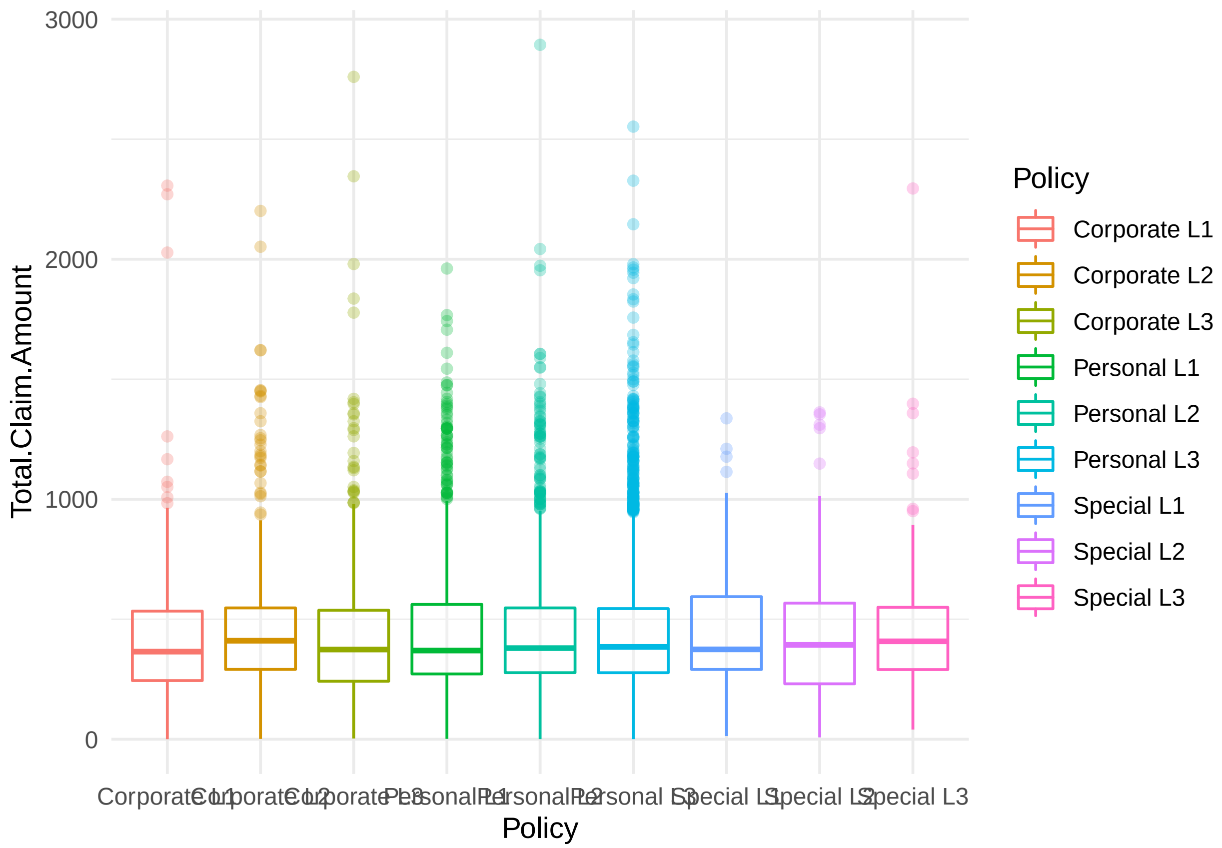
Marital status and total claim. Single people have a higher total claim on average.

```
data %>%
  ggplot(aes(x=reorder(Marital.Status, -Total.Claim.Amount),y=Total.Claim.Amount))+
  geom_boxplot(aes(color=Marital.Status), alpha=0.3)+
  xlab('')
```



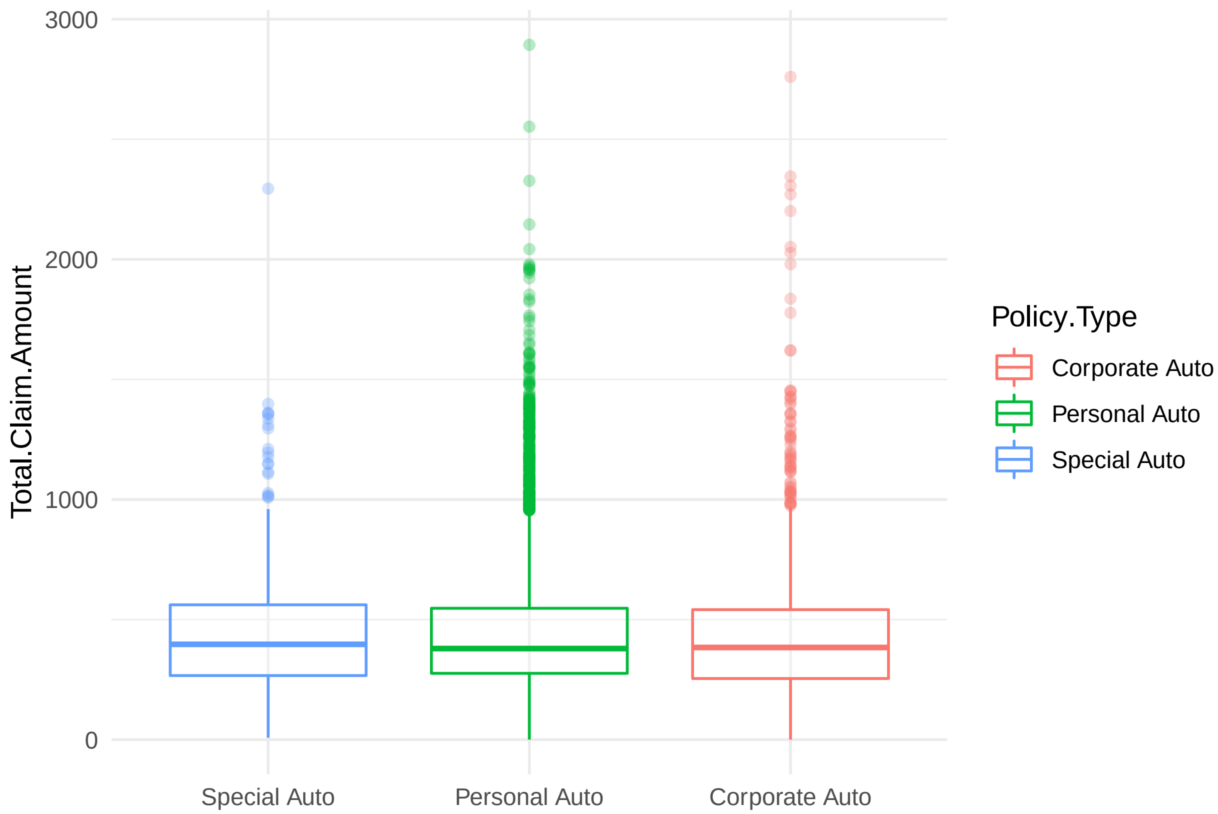
Policies and total claim.

```
data %>%
  ggplot(aes(x=Policy,y=Total.Claim.Amount))+
  geom_boxplot(aes(color=Policy), alpha=0.3)
```



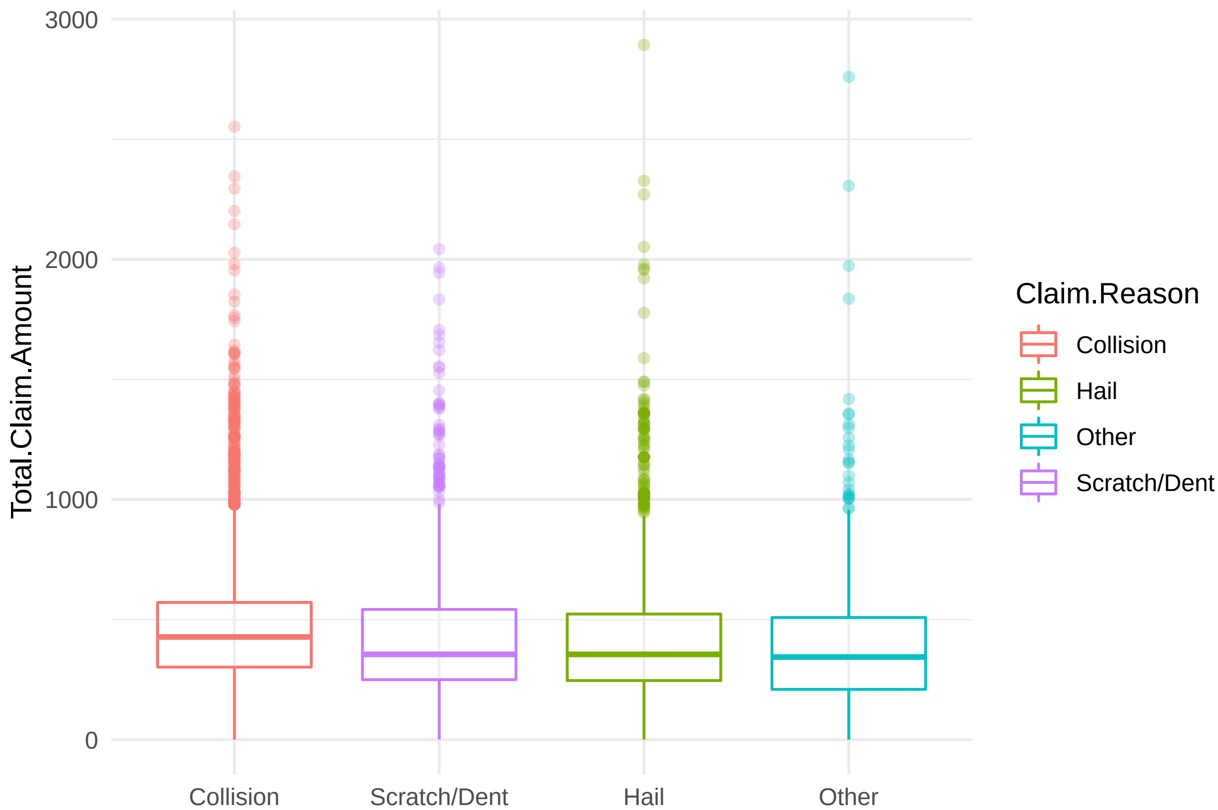
Policy type and total claim.

```
data %>%
  ggplot(aes(x=reorder(Policy.Type, -Total.Claim.Amount),y=Total.Claim.Amount))+
  geom_boxplot(aes(color=Policy.Type), alpha=0.3)+
  xlab('')
```



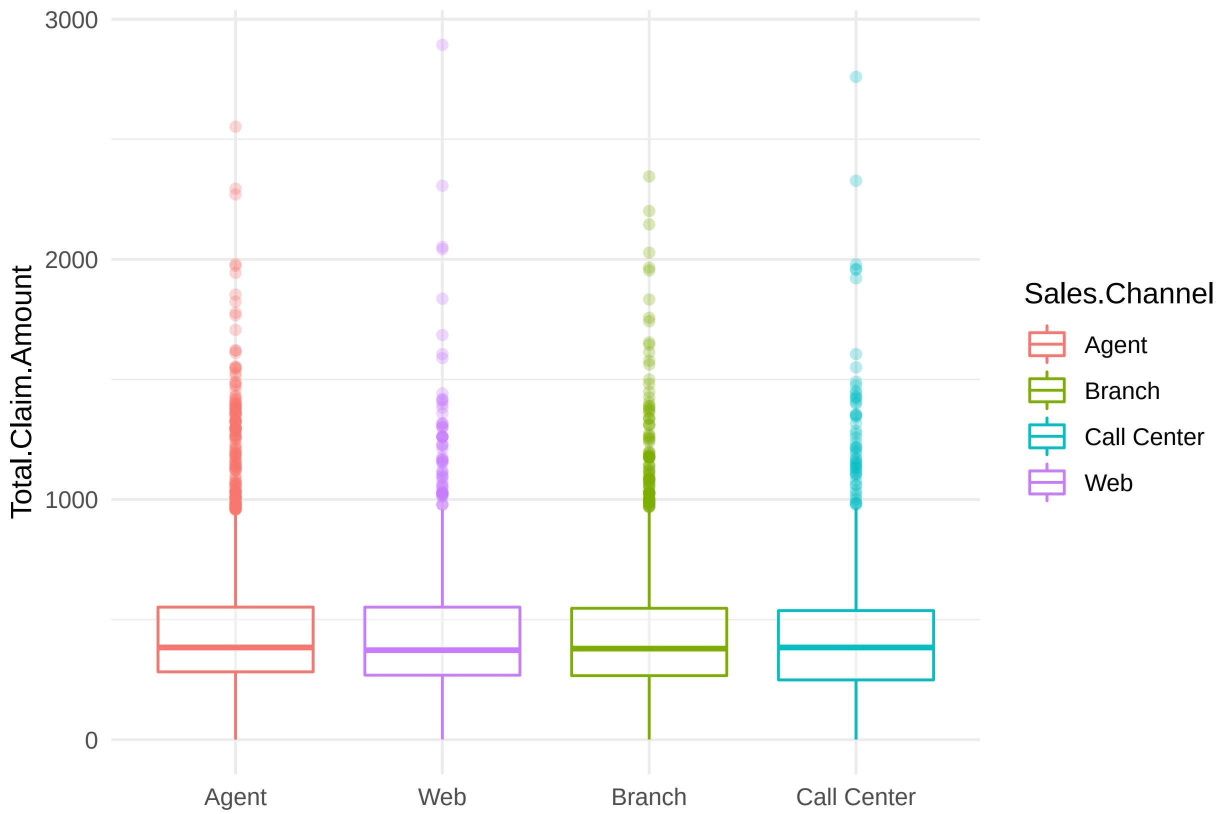
Claim reason and total claim.

```
data %>%
  ggplot(aes(x=reorder(Claim.Reason, -Total.Claim.Amount),y=Total.Claim.Amount))+
  geom_boxplot(aes(color=Claim.Reason), alpha=0.3)+
  xlab('')
```



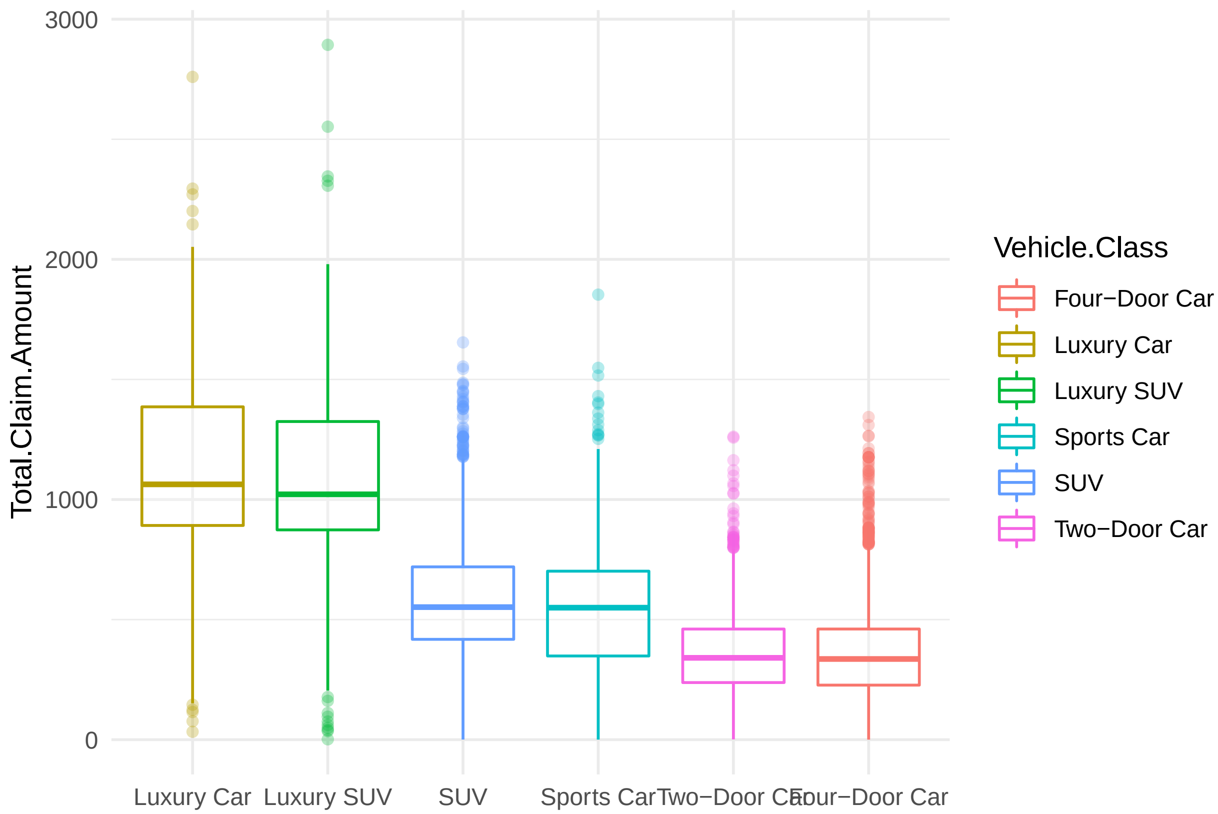
Sales channel and total claim.

```
data %>%
  ggplot(aes(x=reorder(Sales.Channel, -Total.Claim.Amount),y=Total.Claim.Amount))+
  geom_boxplot(aes(color=Sales.Channel), alpha=0.3)+
  xlab('')
```



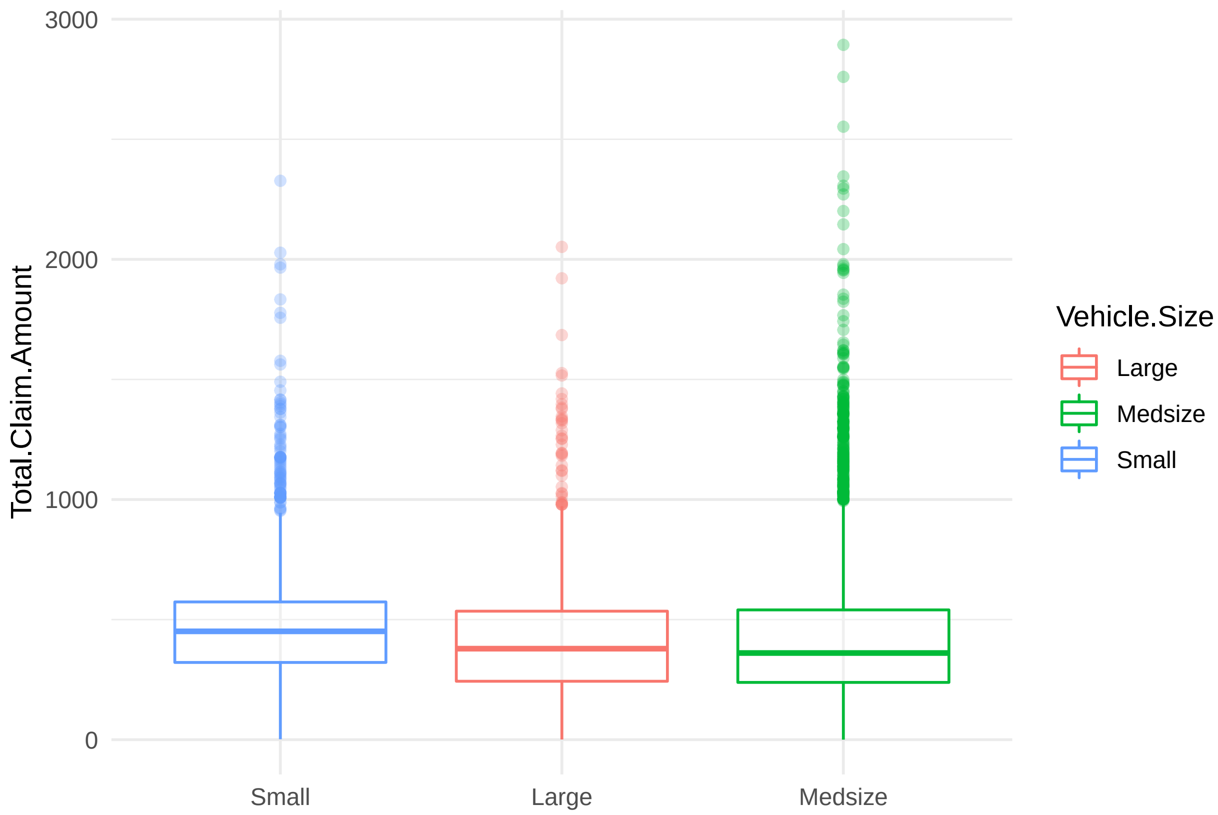
Vehicle class and total claim. Luxury cars have a higher claim on average.

```
data %>%
  ggplot(aes(x=reorder(Vehicle.Class, -Total.Claim.Amount),y=Total.Claim.Amount))+
  geom_boxplot(aes(color=Vehicle.Class), alpha=0.3)+
  xlab('')
```



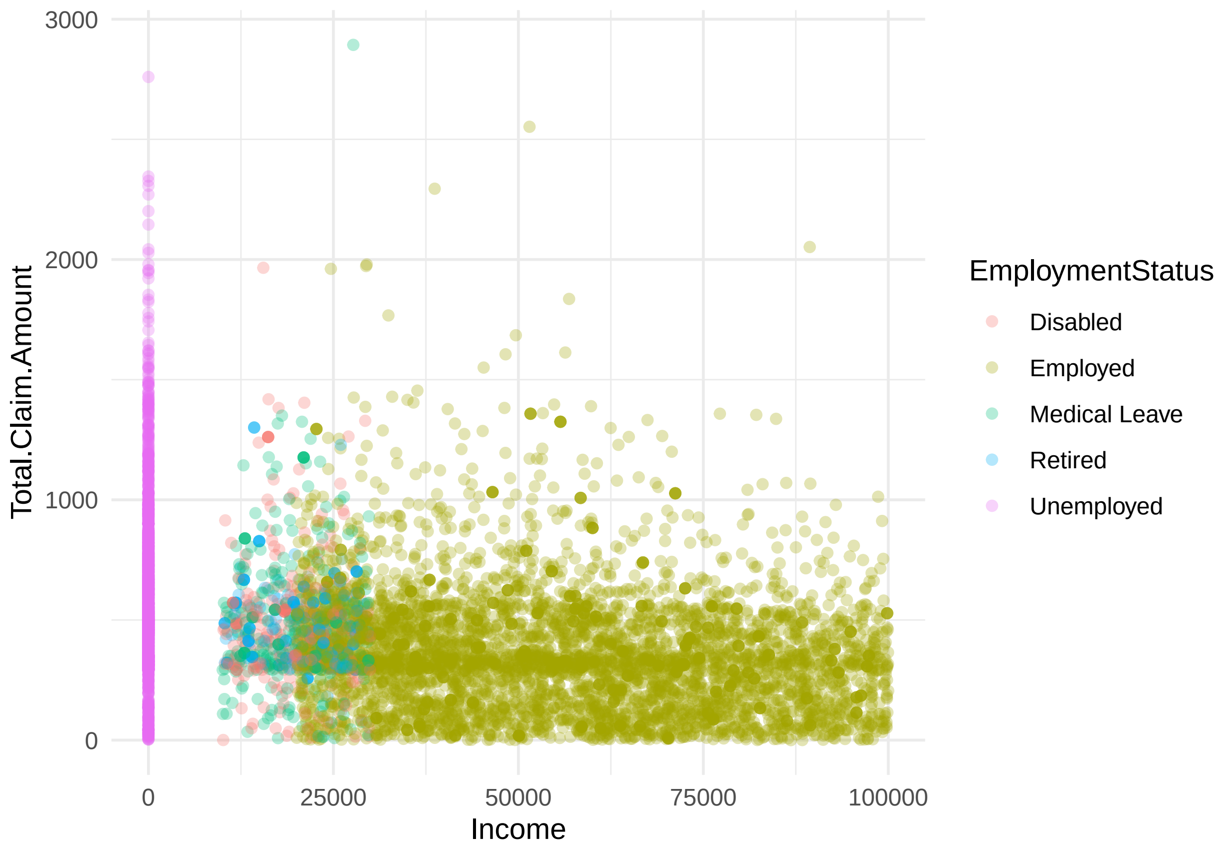
Vehicle size and total claim.

```
data %>%
  ggplot(aes(x=reorder(Vehicle.Size, -Total.Claim.Amount),y=Total.Claim.Amount))+
  geom_boxplot(aes(color=Vehicle.Size), alpha=0.3)+
  xlab('')
```



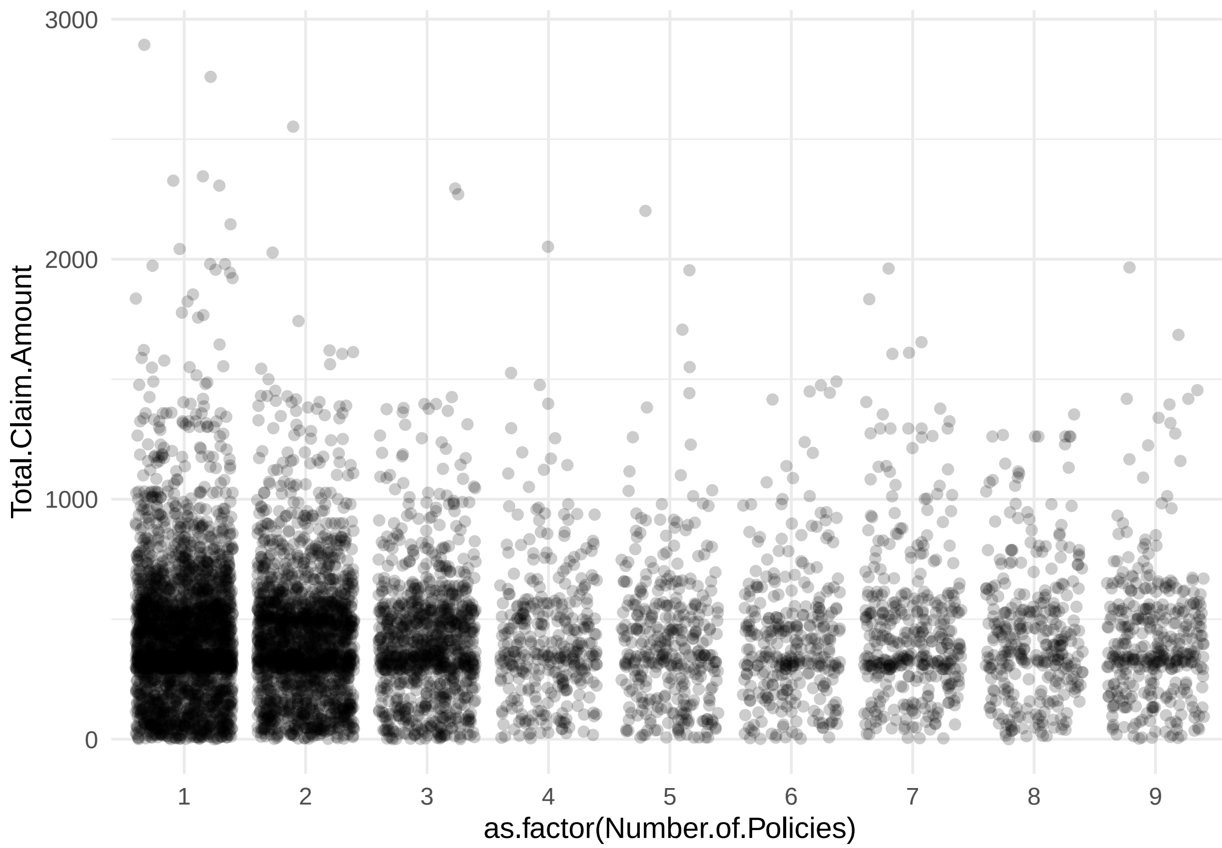
Income and total claim.

```r
data %>%
  ggplot(aes(x=Income,y=Total.Claim.Amount))+
  geom_point(aes(color=EmploymentStatus), alpha=0.3)
```



Total claim by number of policies.

```
data %>%
  ggplot(aes(x=as.factor(Number.of.Policies),y=Total.Claim.Amount))+
  geom_jitter( alpha=0.2)
```


3000
2000
1000
0
Total.Claim.Amount
1
2
3
4
5
6
7
8
9
as.factor(Number.of.Policies)