# Lab1

*David Hou, Jooyeon Seo, Sophia Huang*

## Introduction

A research community studied governments' open access policies: how lowering prices of networks influences broadband market health. While it was revealed that U.S. is among the few countries that did not adopt open access policies out of all OECD countries, network owners insist that attempts to regulate prices or mandate greater penetration will reduce their incentives to invest in modern equipment, and thus, there is a trade-off among price, speed, and penetration.

We are analyzing 3 variables – house penetration, network speed, and price – to see if there are positive/negative relationships among them. In order to find out if there are benefits of open access policies, we're going to run analyses based on network owners' claims. More explicitly:

1. Is the network Speed lower where price is lower? Positive relationship?
2. Is the network Speed lower where penetration is forced high? Negative relationship?
3. Is the network penetration rate lower where price is lower? Positive relationship?

### Setup

First, we load some libraries and the data into R. We do this using dplyr to take advantage of its table manipulation methods. For easy access, we join the three datasets together. A simple full_join is sufficient.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(car)
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(ggplot2)

t1 = as_tibble(read.csv('Penetration.csv', stringsAsFactors = F))
t2 = as_tibble(read.csv('Price.csv', stringsAsFactors = F))
t3 = as_tibble(read.csv('Speed.csv', stringsAsFactors = F))

t = t1 %>% full_join(t2) %>% full_join(t3)
```

```
## Joining, by = c("Country", "Country.Code")
```

```
## Joining, by = "Country"
```

```
str(t)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    32 obs. of  33 variables:
##  $ Country                                   : chr  "Australia" "Austria" "Belgium" "Canada" ...
##  $ Country.Code                              : chr  "AU" "AT" "BE" "CA" ...
##  $ Penetration.per.100.OECD..2008            : num  25.4 21.6 28.1 29 17.2 ...
##  $ Penetration.per.100.OECD..2007            : num  22.8 19.6 25.8 27.2 14.6 ...
##  $ Household.penetration..OECD               : num  52 46.1 56.4 64.2 28.1 ...
##  $ X2G.and.3G.penetration.per.100..OECD      : num  102.1 118.5 96.3 62.1 127.3 ...
##  $ Penetration.per.100.GC                    : num  28.6 20.9 29.1 26.8 15.7 36.9 31.1 27.5 26.7 13
##  $ X3G.penetration.per.100                   : num  55.26 32.37 20.05 6.61 10.19 ...
##  $ Growth.in.3G.penetration                  : chr  "510.90%" "339.30%" "239.70%" "225.50%" ...
##  $ Wi.Fi.hotspots..JiWire                    : int  2611 986 2318 3576 429 1206 750 25625 14512 531
##  $ Wi.Fi.hotspots.per.100.000..JiWire        : num  12.26 12.02 22.29 10.67 4.21 ...
##  $ Percent.of.population.in.urban.areas      : chr  "88%" "66%" "97%" "80%" ...
##  $ X                                         : logi  NA NA NA NA NA NA ...
##  $ Price.for.low.speeds..combined            : chr  " $51.86 " " $31.55 " "" " $28.02 " ...
##  $ Price.for.med.speeds..combined            : chr  " $61.02 " " $30.76 " " $31.61 " " $40.48 " ...
##  $ Price.for.high.speeds..combined           : chr  " $60.19 " " $52.54 " " $50.41 " " $63.77 " ...
##  $ Price.for.very.high.speeds..combined      : chr  "" " $80.07" "" " $121.94" ...
##  $ Country.code                              : chr  "AU" "AT" "BE" "CA" ...
##  $ Maximum.advertised.speed.OECD..kbps.      : chr  " 30,000 " " 25,600 " " 20,000 " " 25,000 " ...
##  $ Average.advertised.speed.OECD..kbps.      : chr  " 15,539 " " 10,292 " " 7,544 " " 6,236 " ...
##  $ Average.actual.speed..Akamai..kbps.       : chr  " 2,499 " " 3,773 " " 4,737 " " 3,786 " ...
##  $ Average.download.speedtest.net..kbps.     : chr  " 4,602 " " 5,683 " " 6,908 " " 4,783 " ...
##  $ Standard.deviation.download..speedtest.net: chr  " 6,504 " " 10,049 " " 9,754 " " 6,528 " ...
##  $ Average.upload.speedtest.net..kbps.       : chr  " 641 " " 973 " " 708 " " 876 " ...
##  $ Standard.deviation.upload..speedtest.net  : chr  " 2,298 " " 2,847 " " 2,762 " " 2,940 " ...
##  $ Average.latency.speedtest.net             : num  162 115 87 127 102 100 130 145 123 133 ...
##  $ Standard.deviation.latency..speedtest.net : chr  " 370 " " 271 " " 246 " " 309 " ...
##  $ Median.download..speedtest.net..kbps.     : chr  " 2,550 " " 3,159 " " 3,954 " " 3,399 " ...
##  $ Median.upload..speedtest.net..kbps.       : chr  " 345 " " 455 " " 394 " " 531 " ...
##  $ Median.latency..speedtest.net             : int  68 53 42 62 39 43 48 89 61 68 ...
##  $ X90p..Download..speedtest.net..kbps.      : chr  " 10,676 " " 12,338 " " 14,911 " " 9,356 " ...
##  $ X90p..Upload..speedtest.net..kbps.        : chr  " 853 " " 1,467 " " 922 " " 960 " ...
##  $ X10p..Latency..speedtest.net              : int  22 18 18 19 10 11 14 34 21 26 ...
```

## Cleaning the Data

We see a couple of issues with the data:

1. There are two blank rows at the bottom, which We filter out of the table.

2. The 'X' variable seems to be empty, so we omit that column.
3. Due to a typo in the original data, Both 'Country.Code' and 'Country.code' exist. We omit the latter.
4. There are percent signs, dollar signs, and commas being used as decimal separators. There's probably some R function better than read.csv for cleaning these columns, but let's use some regular expressions just for fun.

```
clean_percent = function(x) as.numeric(gsub('%', '', x))
clean_dollar = function(x) as.numeric(gsub('\\$', '', x))
clean_commas = function(x) as.numeric(gsub(',', '', x))
```

```
t = t %>%
    filter(Country != '') %>%
    select(-c(X, Country.code)) %>%
    mutate(Growth.in.3G.penetration = clean_percent(Growth.in.3G.penetration)) %>%
    mutate(Percent.of.population.in.urban.areas = clean_percent(Percent.of.population.in.urban.areas))

for (c in grep('Price', names(t))) t[c] = unlist(lapply(t[c], clean_dollar))
for (c in grep('Maximum|Average|Standard|Median|X90p', names(t))) t[c] = unlist(lapply(t[c], clean_comma
t$Country = factor(t$Country)
t$Country.Code = factor(t$Country.Code)
str(t)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    30 obs. of  31 variables:
##  $ Country                                   : Factor w/ 30 levels "Australia","Austria",..: 1 2 3 4
##  $ Country.Code                              : Factor w/ 30 levels "AT","AU","BE",..: 2 1 3 4 6 8 10
##  $ Penetration.per.100.OECD..2008            : num  25.4 21.6 28.1 29 17.2 ...
##  $ Penetration.per.100.OECD..2007            : num  22.8 19.6 25.8 27.2 14.6 ...
##  $ Household.penetration..OECD               : num  52 46.1 56.4 64.2 28.1 ...
##  $ X2G.and.3G.penetration.per.100..OECD      : num  102.1 118.5 96.3 62.1 127.3 ...
##  $ Penetration.per.100.GC                    : num  28.6 20.9 29.1 26.8 15.7 36.9 31.1 27.5 26.7 13
##  $ X3G.penetration.per.100                   : num  55.26 32.37 20.05 6.61 10.19 ...
##  $ Growth.in.3G.penetration                  : num  511 339 240 226 144 ...
##  $ Wi.Fi.hotspots..JiWire                    : int  2611 986 2318 3576 429 1206 750 25625 14512 531
##  $ Wi.Fi.hotspots.per.100.000..JiWire        : num  12.26 12.02 22.29 10.67 4.21 ...
##  $ Percent.of.population.in.urban.areas      : num  88 66 97 80 74 86 61 77 75 59 ...
##  $ Price.for.low.speeds..combined            : num  51.9 31.6 NA 28 32 ...
##  $ Price.for.med.speeds..combined            : num  61 30.8 31.6 40.5 43.4 ...
##  $ Price.for.high.speeds..combined           : num  60.2 52.5 50.4 63.8 68.8 ...
##  $ Price.for.very.high.speeds..combined      : num  NA 80.1 NA 121.9 105.9 ...
##  $ Maximum.advertised.speed.OECD..kbps.      : num  30000 25600 20000 25000 20480 ...
##  $ Average.advertised.speed.OECD..kbps.      : num  15539 10292 7544 6236 10468 ...
##  $ Average.actual.speed..Akamai..kbps.       : num  2499 3773 4737 3786 4381 ...
##  $ Average.download.speedtest.net..kbps.     : num  4602 5683 6908 4783 5776 ...
##  $ Standard.deviation.download..speedtest.net: num  6504 10049 9754 6528 8244 ...
##  $ Average.upload.speedtest.net..kbps.       : num  641 973 708 876 1847 ...
##  $ Standard.deviation.upload..speedtest.net  : num  2298 2847 2762 2940 4901 ...
##  $ Average.latency.speedtest.net             : num  162 115 87 127 102 100 130 145 123 133 ...
##  $ Standard.deviation.latency..speedtest.net : num  370 271 246 309 205 217 281 297 251 253 ...
##  $ Median.download..speedtest.net..kbps.     : num  2550 3159 3954 3399 3614 ...
##  $ Median.upload..speedtest.net..kbps.       : num  345 455 394 531 524 777 646 665 592 452 ...
##  $ Median.latency..speedtest.net             : num  68 53 42 62 39 43 48 89 61 68 ...
##  $ X90p..Download..speedtest.net..kbps.      : num  10676 12338 14911 9356 11493 ...
##  $ X90p..Upload..speedtest.net..kbps.        : num  853 1467 922 960 3912 ...
##  $ X10p..Latency..speedtest.net              : int  22 18 18 19 10 11 14 34 21 26 ...
```

Now we have a tidy data set. All the columns hold either numeric or integer data with the exceptions of 'Country' and 'Country.Code'.

## Univariate Analysis of Key Variables

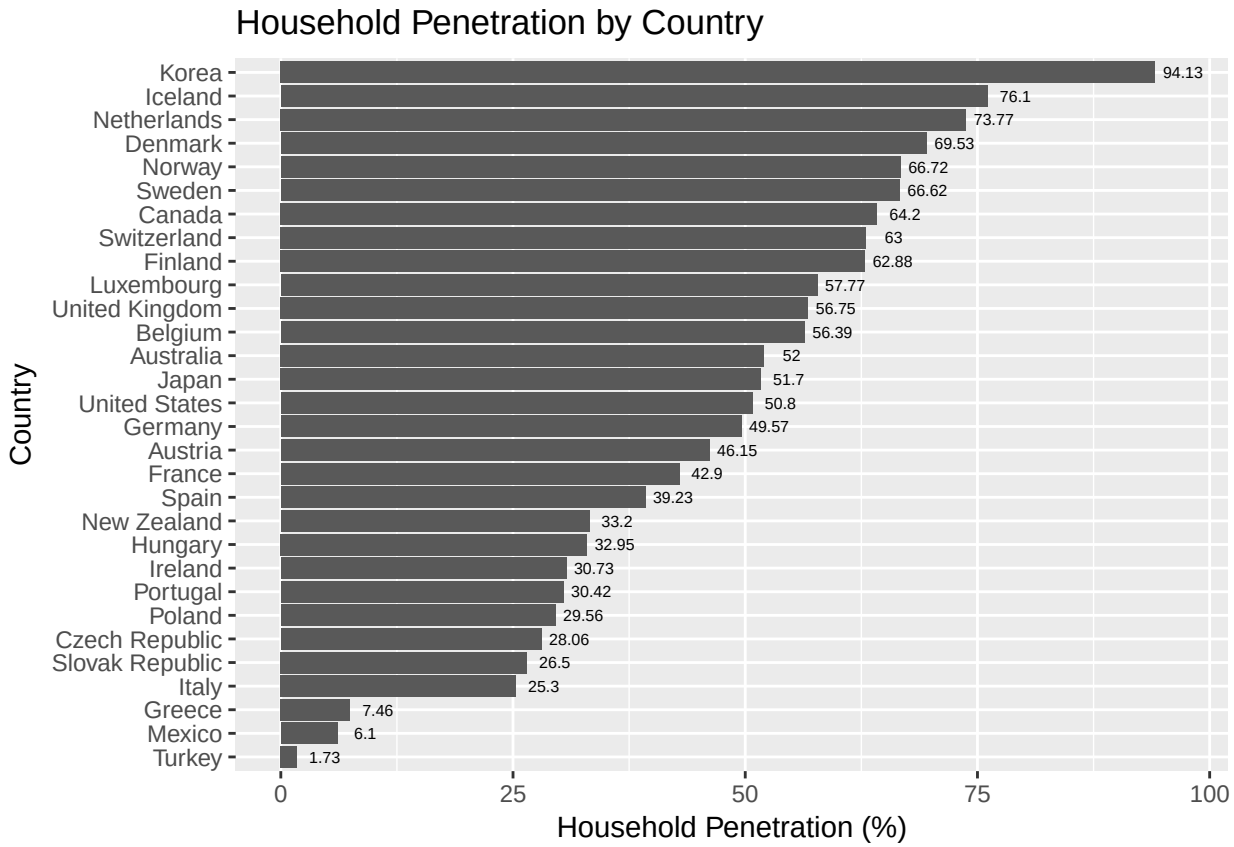
For looking at key variables by country, we will use ggplot since it makes better barplots than the base package.

First let's take a look at household penetration by country.

```
summary(t$Household.penetration..OECD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.73   30.50   50.19   46.41   62.97   94.13
```

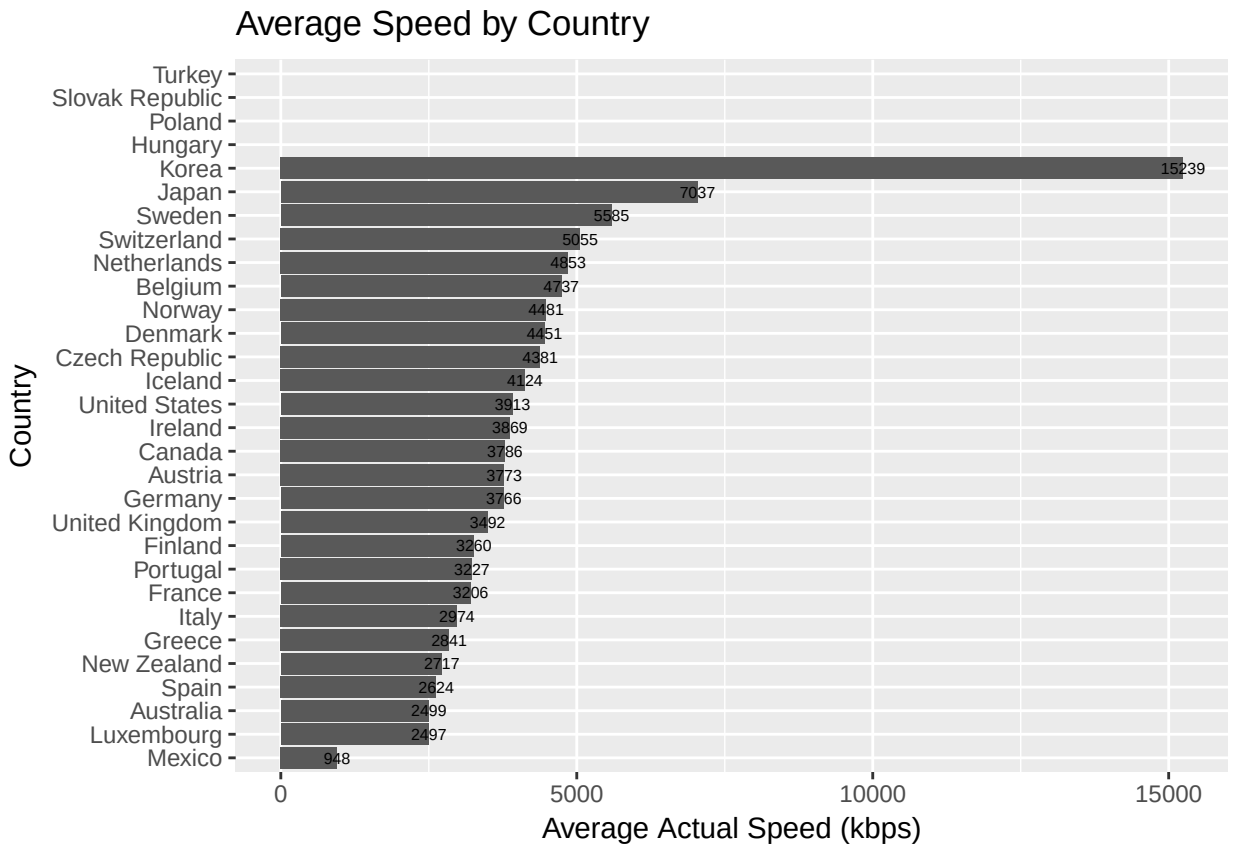
```
ggplot(t, aes(x=reorder(Country, Household.penetration..OECD), y=Household.penetration..OECD)) +
    geom_bar(stat='identity') +
    geom_text(aes(label=Household.penetration..OECD), nudge_y = 3, size = 2) +
    labs(title = 'Household Penetration by Country', y = 'Household Penetration (%)', x = 'Country') +
    coord_flip()
```




Korea has the highest Household penetration with 94.13%. The United States is at median with 50.8%. Turkey has the lowest penetration at 1.73%.

Let's also take a look at average speed.

```
ggplot(t, aes(x=reorder(Country, Average.actual.speed..Akamai..kbps.), y=Average.actual.speed..Akamai..
    geom_bar(stat='identity') +
    geom_text(aes(label=Average.actual.speed..Akamai..kbps.), nudge_y = 3, size = 2) +
    labs(title = 'Average Speed by Country', y = 'Average Actual Speed (kbps)', x = 'Country') +
    coord_flip()
```




We immediately notice a few countries with missing data. It is worth noting that the countries with missing data (Turkey, Slovak Republic, Poland, and Hungary) were also very low in household penetration. Here we see that Korea, which had the highest penetration, also has the highest speed at 15239 kbps. In fact Korea's internet is over twice as fast as the second best (Japan). We will see this later, as Korea shows up as an outlier in many analyses.

We also see that United States has a medium speed at 3913 kbps. Mexico has lowest speed at 948 kbps.

Finally, let's take a look at some prices. It is interesting to consider why certain countries are missing prices at certain speed tiers. Is it a data collection error or do they simply not provide those speeds? Note that we do not actually know where the cut off points for these speed tiers are in the original dataset (e.g. very high vs. high).

```
s1 = t %>% select(c(Country, Price.for.very.high.speeds..combined)) %>% mutate(speed = 'Very High') %>%
    rename(price = Price.for.very.high.speeds..combined)
s2 = t %>% select(c(Country, Price.for.high.speeds..combined)) %>% mutate(speed = 'High') %>% rename(pri
s3 = t %>% select(c(Country, Price.for.med.speeds..combined)) %>% mutate(speed = 'Medium') %>% rename(p
s4 = t %>% select(c(Country, Price.for.low.speeds..combined)) %>% mutate(speed = 'Low') %>% rename(price
s = s1 %>% bind_rows(s2) %>% bind_rows(s3) %>% bind_rows(s4) %>% mutate(speed = factor(speed, c('Very Hi

ggplot(s, aes(x = Country, y = price)) +
    geom_bar(stat='identity') +
    facet_grid(. ~ speed) +
    labs(title = 'Price by Country', y = 'Price', x = 'Country') +
    coord_flip()
```


This visual shows quite a lot of missing data for very high prices. Other interesting things of note is abnormally high price in Poland (which is missing pricing data altogether for very high speeds). We see that Korea, which has great penetration and speeds, also has very reasonable pricing.

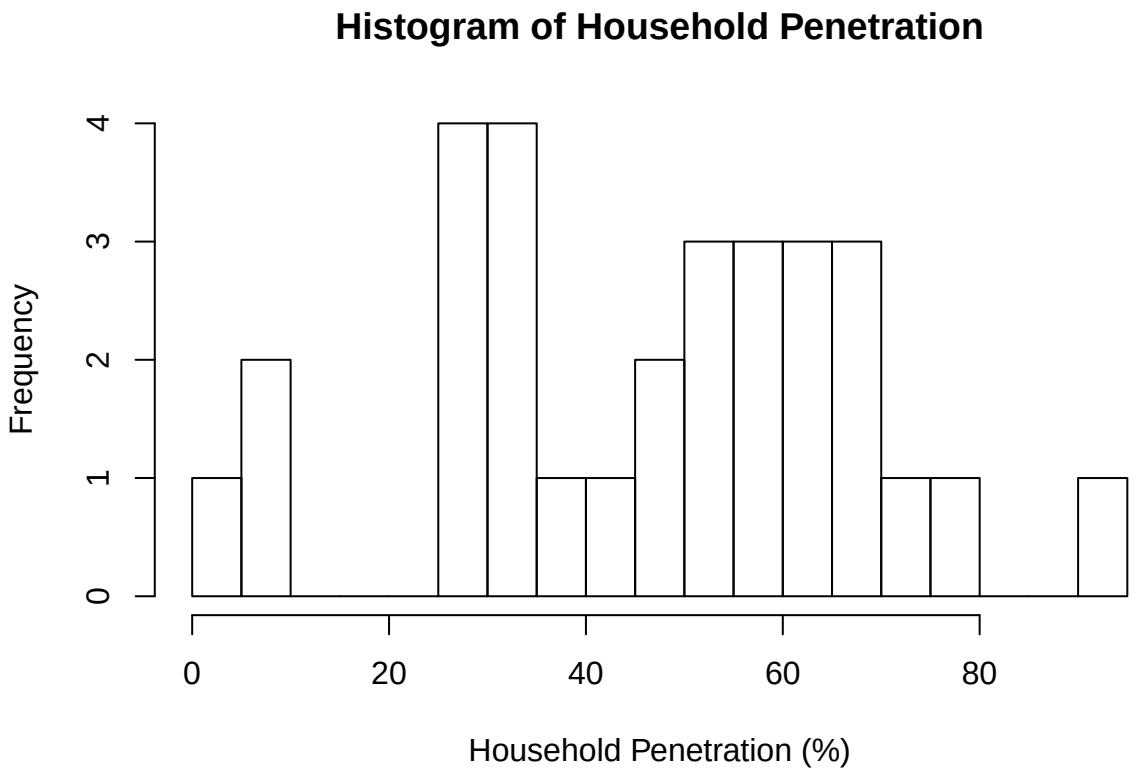
Now we examine the overall distribution of key variables - actual network speed, household penetration, and price for different speeds, with no segregation among countries.

## Household penetration

```
summary(t$Household.penetration..OECD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.73   30.50   50.19   46.41   62.97   94.13
```

```
hist(t$Household.penetration..OECD, breaks = 20,
     main = 'Histogram of Household Penetration', xlab = 'Household Penetration (%)')
```




We see that most of the countries' household penetration lies between 25% and 80%.

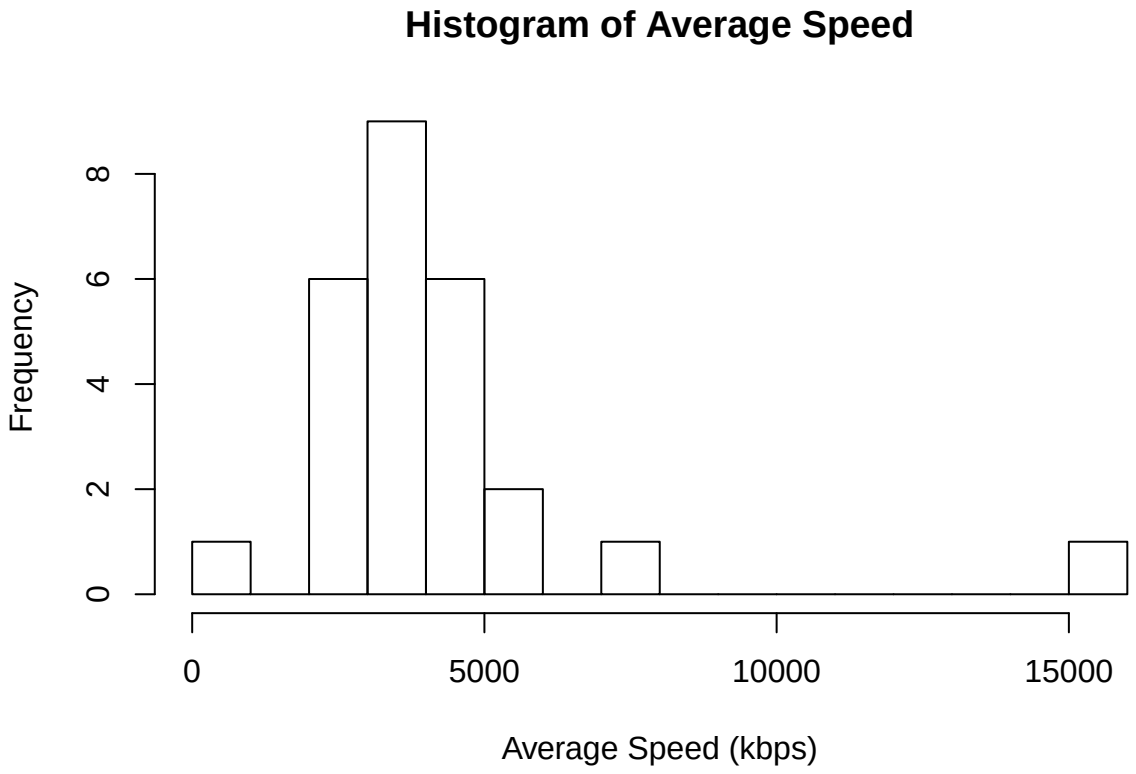
### Average Network Speed

The plotted speeds are measured by Akamai.

```
summary(t$Average.actual.speed..Akamai..kbps., na.rm = T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     948    3032    3780    4205    4474   15239       4
```

```
hist(t$Average.actual.speed..Akamai..kbps., breaks = 20,
     main = 'Histogram of Average Speed', xlab = 'Average Speed (kbps)')
```

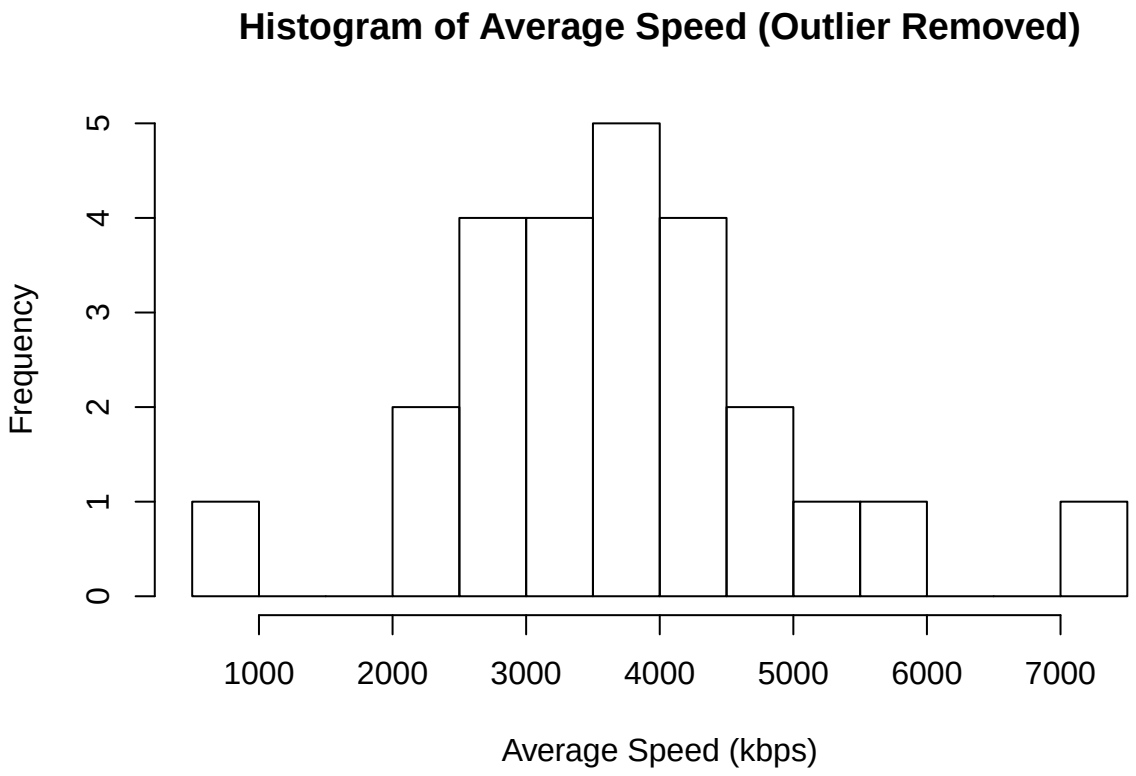



There seems to be an outlier at high end (Korea with its ridiculously good internet), which makes it difficult to examine the distribution of the values where values are clustered. We remove the outlier for now, so we can get a better look at the rest of the countries.

```
t_sub1 <- subset(t, Average.actual.speed..Akamai..kbps. < 8000, na.rm = T)
summary(t_sub1$Average.actual.speed..Akamai..kbps.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     948    2974    3773    3764    4451    7037
```

```
hist(t_sub1$Average.actual.speed..Akamai..kbps., breaks = 20,
     main = 'Histogram of Average Speed (Outlier Removed)', xlab = 'Average Speed (kbps)')
```




We now see that mean and median value are fairly close for these countries (low skewness). Most countries have network speed between 1000 and 7000. The distribution of speed actually looks fairly normal.

## Price for various speeds

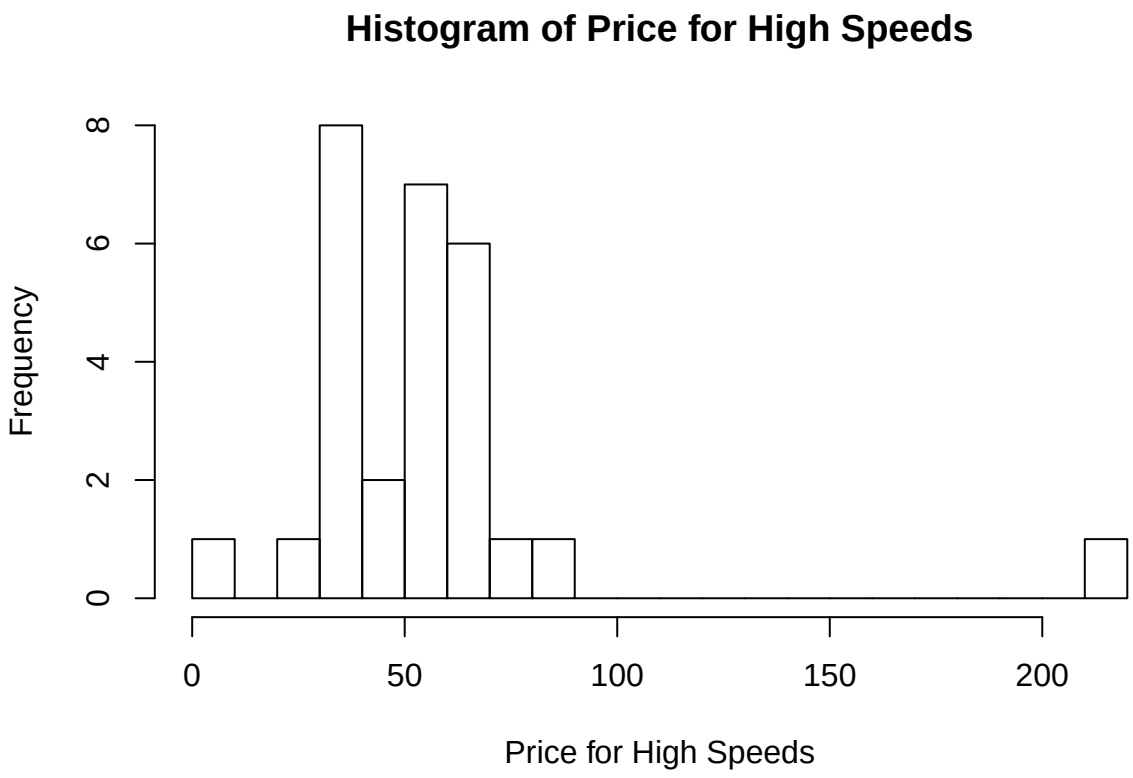
Since Very-high-speed price lacks many data points, and since Low-speed price and Med-speed price show similar distribution, we'll examine High-speed price and Med-speed price for further analyses.

### Price for high speed network

```
summary(t$Price.for.high.speeds..combined, na.rm = T)
```

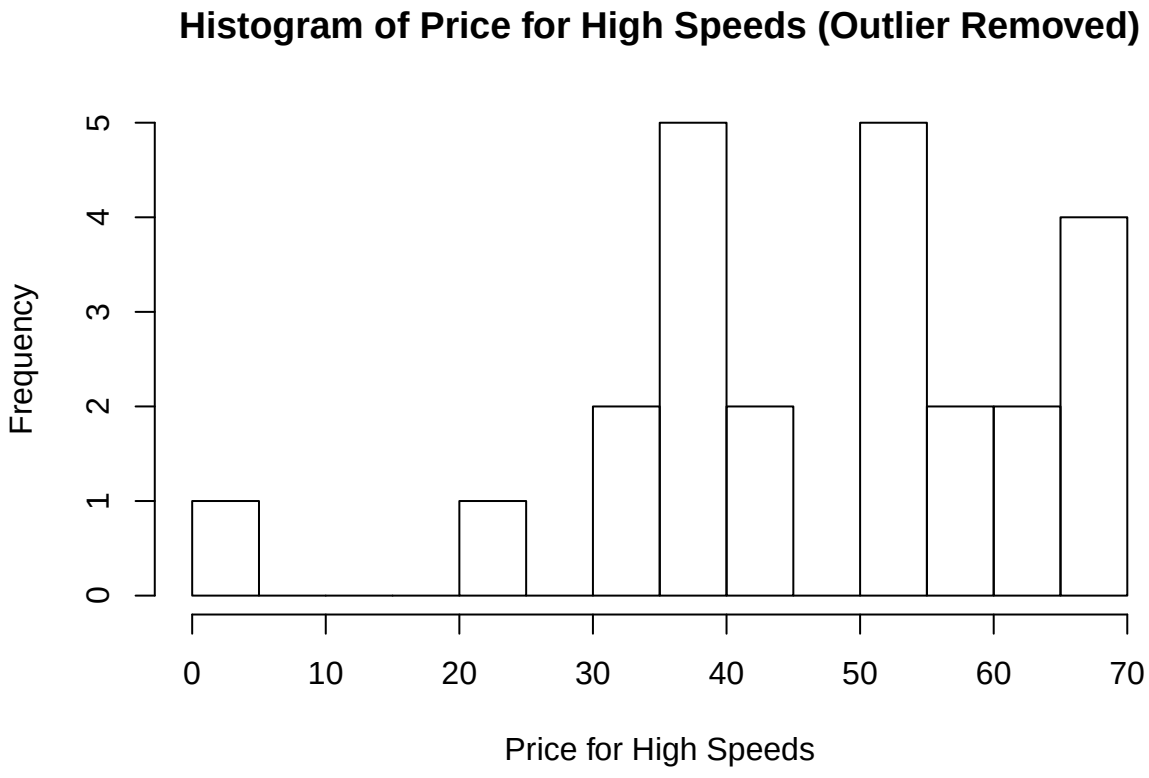
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##   0.6931  38.1225  53.1600  55.6437  64.2075 210.3600        2
```

```
hist(t$Price.for.high.speeds..combined, breaks = 20,
     main = 'Histogram of Price for High Speeds', xlab = 'Price for High Speeds')
```




We can see that the outlier (Poland) above 200 makes examination of distribution difficult. Let us take a look at the distribution sans that one data point.

```
t_sub2 <- subset(t_sub1, Price.for.high.speeds..combined < 200, na.rm = T)
hist(t_sub2$Price.for.high.speeds..combined, breaks = 20,
     main = 'Histogram of Price for High Speeds (Outlier Removed)', xlab = 'Price for High Speeds')
```




From

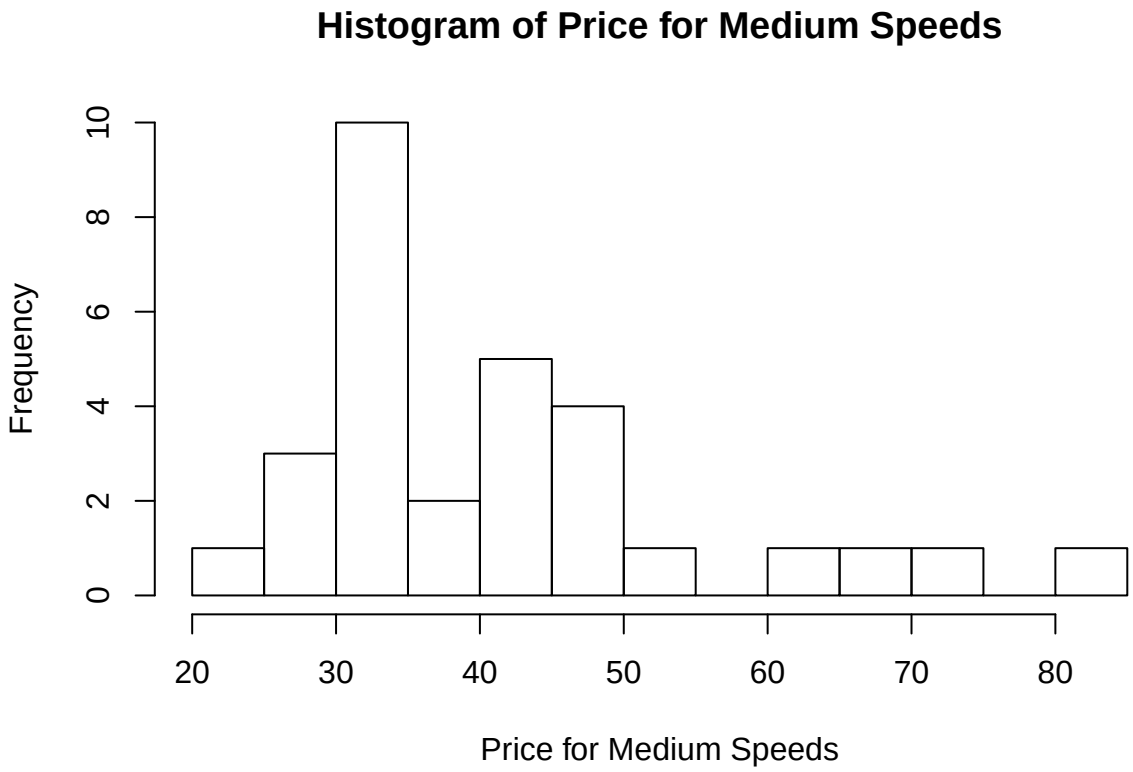
this revised subset, we can see that most countries charge 30 to 70 dollars for high speed internet varying diversely.

**Price for Med speed network**

```
summary(t$Price.for.med.speeds..combined)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.32   31.62   37.32   41.45   45.71   82.76
```

```
hist(t$Price.for.med.speeds..combined, breaks = 20,
     main = 'Histogram of Price for Medium Speeds', xlab = 'Price for Medium Speeds')
```




Compared to high speed networks, medium speed network shows more clustering, around 30-50 dollars.

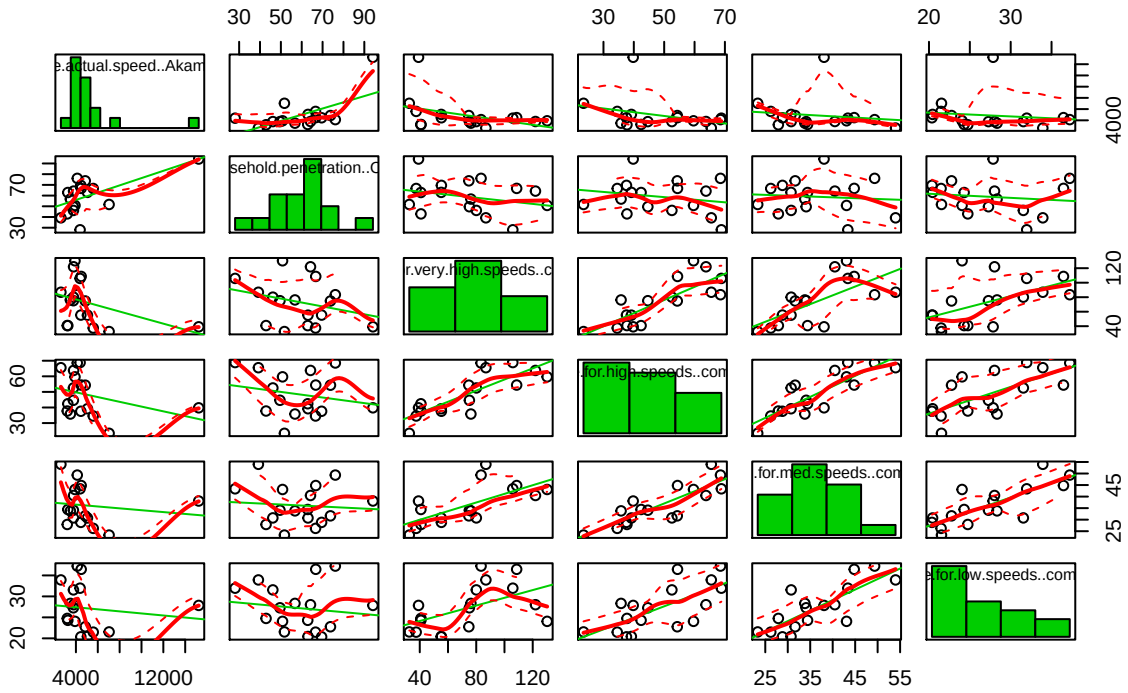
# Analysis of Key Relationships

## Scatterplot Matrix

To examine relationships among the key variables, we first draw some scatterplot matrices.

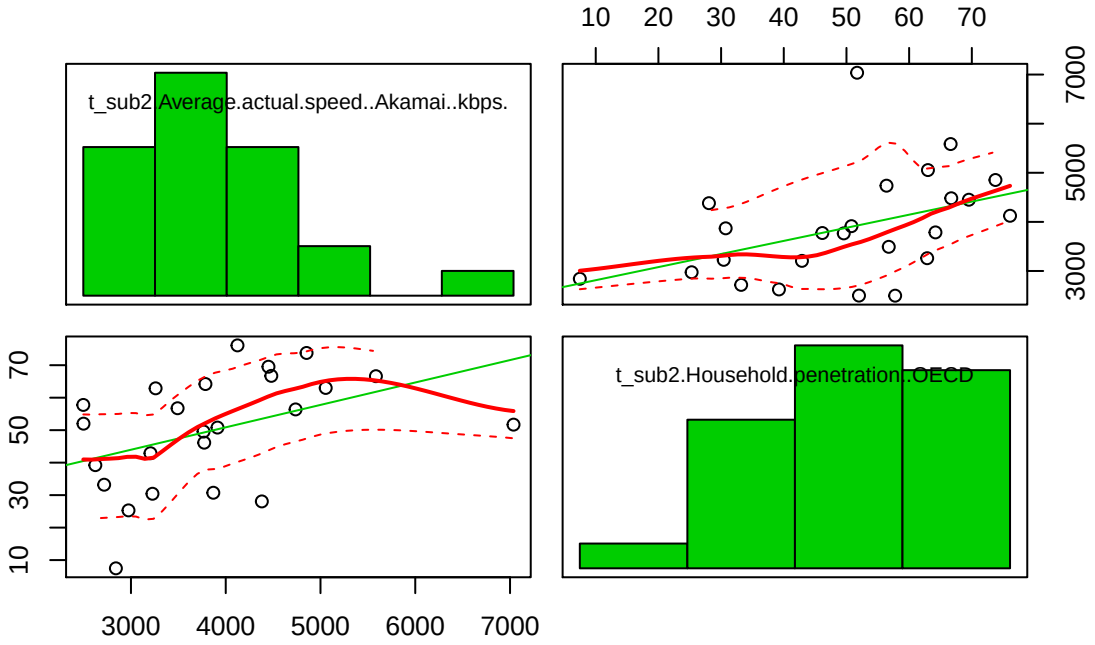
```
scatterplotMatrix( ~t$Average.actual.speed..Akamai..kbps. +
                     t$Household.penetration..OECD +
                     t$Price.for.very.high.speeds..combined +
                     t$Price.for.high.speeds..combined +
                     t$Price.for.med.speeds..combined +
                     t$Price.for.low.speeds..combined,
```

```
                  diagonal = 'hist',
                  main = 'Average Speed vs Household Penetration vs Price')
```




The first column of this scatterplotMatrix is one of our main interests, which shows the relationship between 1) speed and household penetration, and 2) speed and price for speed tiers. However, we can instantly notice that these scatterplots are affcted heavily by one outlier at the far end (Korea). Let's remove the outlier for now and separate each relationship to see them more clearly.

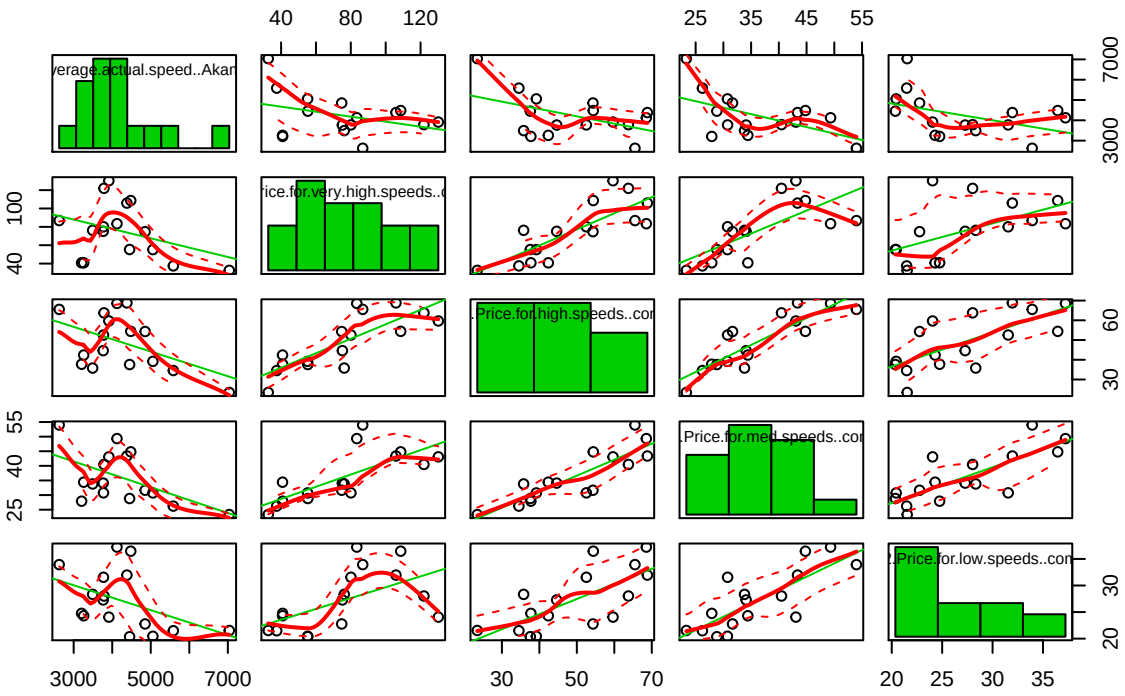
```
scatterplotMatrix( ~ t_sub2$Average.actual.speed..Akamai..kbps. +
                     t_sub2$Household.penetration..OECD,
                  diagonal = 'hist',
                  main = 'Average Speed vs Houshold Penetration')
```




We see a slightly positive relationship between speed and penetration. Countries with higher household penetration also have higher speeds.

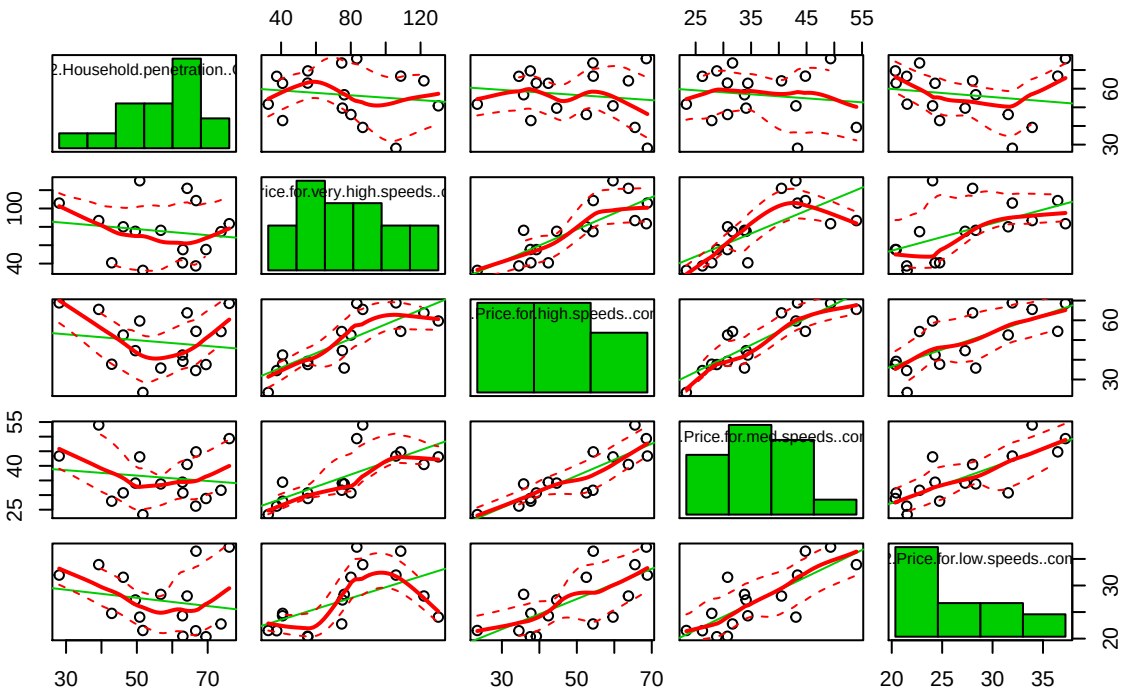
Next, let's draw a scatterplot matrix for speed against the various price points.

```
scatterplotMatrix( ~ t_sub2$Average.actual.speed..Akamai..kbps. +
                     t_sub2$Price.for.very.high.speeds..combined +
                     t_sub2$Price.for.high.speeds..combined +
                     t_sub2$Price.for.med.speeds..combined +
                     t_sub2$Price.for.low.speeds..combined,
                   diagonal = 'hist',
                   main = 'Average Speed vs Prices for various speed tiers')
```




Perhaps uninterestingly, the prices at various speed breakpoints track each other very well. A country with cheap internet at one breakpoint has cheap internet at the others; one with expensive internet has expensive internet at all breakpoints as well. The more interesting relationship is that speed seems to be negatively related to all four tiers of pricing. That is, the faster the internet a country has, the lower the pricing.

```
scatterplotMatrix( ~ t_sub2$Household.penetration..OECD +
                     t_sub2$Price.for.very.high.speeds..combined +
                     t_sub2$Price.for.high.speeds..combined +
                     t_sub2$Price.for.med.speeds..combined +
                     t_sub2$Price.for.low.speeds..combined,
                   diagonal = 'hist',
                   main = 'Household Penetration vs Prices for various speed tiers')
```




Relationship between household penetration and prices for various network speeds is also slightly negative, which seems natural in 2 ways. 1) There is a positive relationship between Speed and Penetration, and a negative relationship between Speed and Price, so negative relationship between Penetration and Price seems plausible. 2) More intuitively, the cheaper the price for network is, the higher penetration rate is. Countries with lower price has higher network penetration.

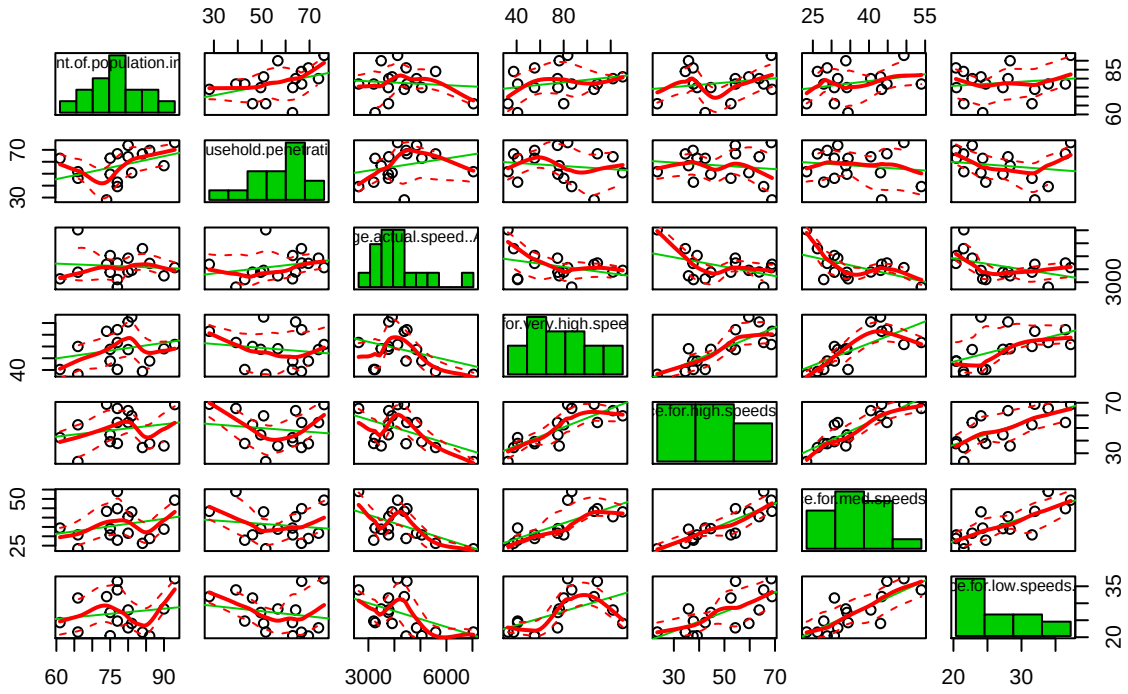
We can clearly see: 1. Positive relationship between Speed and Penetration; 2. Negative relationships between Speed and Prices, more so with Price for med and low speed networks than Price for high speed networks. The reason why there is weaker relationship between Speed and price for high/veryhigh than Speed and price for low/med is because prices for high/veryhigh network speeds are dispersed more broadly than low/med network. 3. Negative relationship between Penetration and Price.

## Analysis of secondary effects

We're now curious to know what other factors might affect Speed, Penetration, and Price of internets. First, we examined the effect of population density in urban areas, to check if the network owners' following claim is valid: attempts to regulate prices or mandate greater penetration will reduce their incentives to lay cable to rural areas.

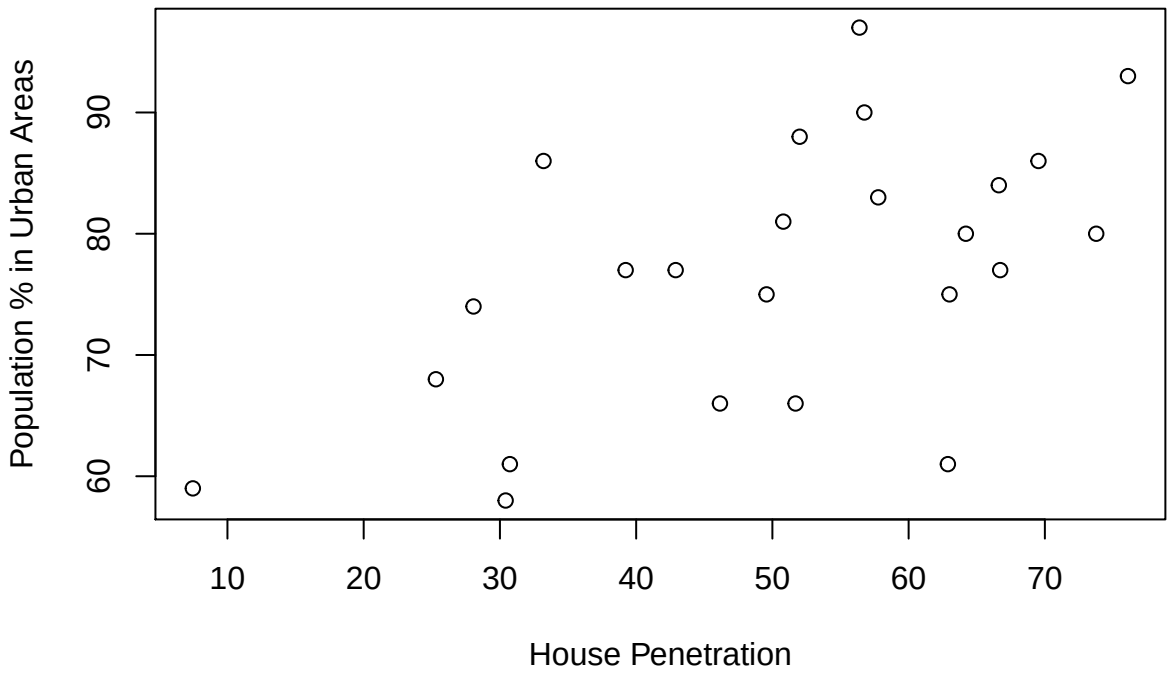
```
scatterplotMatrix( ~ t_sub2$Percent.of.population.in.urban.areas +
                     t_sub2$Household.penetration..OECD +
                     t_sub2$Average.actual.speed..Akamai..kbps. +
                     t_sub2$Price.for.very.high.speeds..combined +
                     t_sub2$Price.for.high.speeds..combined +
                     t_sub2$Price.for.med.speeds..combined +
                     t_sub2$Price.for.low.speeds..combined,
                   diagonal = 'hist',
                   main = 'Household Penetration vs Prices for various speed tiers')
```

## Household Penetration vs Prices for various speed tiers

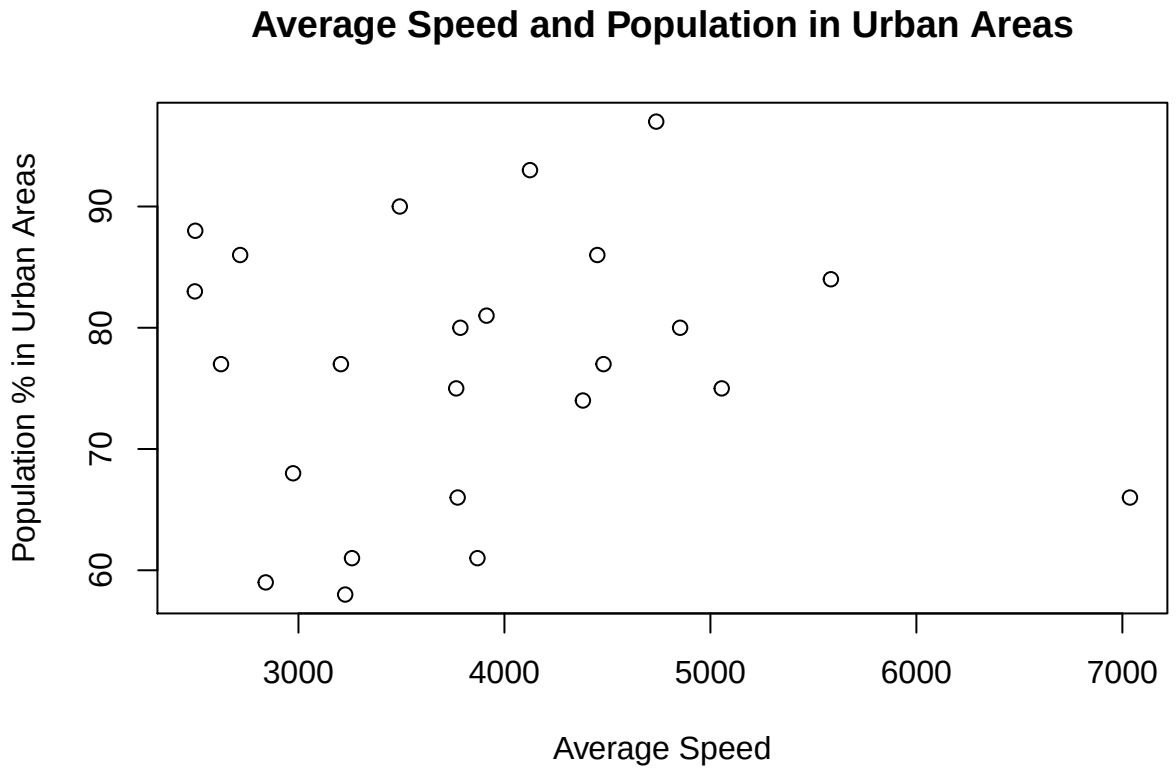



```
plot(t_sub2$Household.penetration..OECD, factor = 2, t_sub2$Percent.of.population.in.urban.areas, facto
     xlab = "House Penetration", ylab = "Population % in Urban Areas",
     main = "House Penetration and Population in Urban Areas")
```

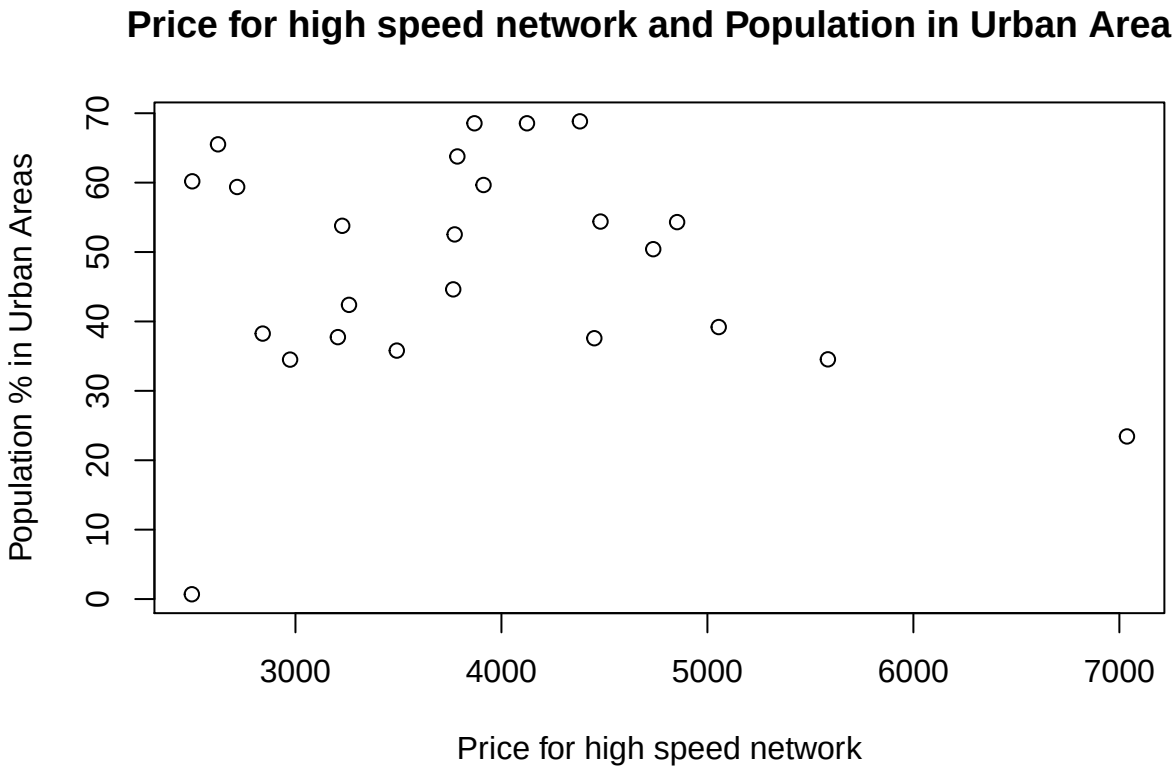
## House Penetration and Population in Urban Areas



```r
plot(t_sub2$Average.actual.speed..Akamai..kbps., factor = 2, t_sub2$Percent.of.population.in.urban.areas
     xlab = "Average Speed", ylab = "Population % in Urban Areas",
     main = "Average Speed and Population in Urban Areas")
```




```r
plot(t_sub2$Average.actual.speed..Akamai..kbps., factor = 2, t_sub2$Price.for.high.speeds..combined, fa
     xlab = "Price for high speed network", ylab = "Population % in Urban Areas",
     main = "Price for high speed network and Population in Urban Areas")
```




From the graphs, we see that there are positive relationships between the Percentage of population in urban areas and House Penetration, and between the Percentage of population in urban areas and Price of various speed tiers. There doesn't seem to be a strong relationship between Percentage of population in urban areas and Speed. We can see that network owners' claim that open access policies lowering price hinders rural penetration is not sound: Countries with more people in the urban area also show higher price.

To further investigate the relationship we use linear regression to see if the average actual speed change according to these two variables.

```
lm(t_sub2$Household.penetration..OECD ~ t_sub2$Percent.of.population.in.urban.areas)
```

```
##
## Call:
## lm(formula = t_sub2$Household.penetration..OECD ~ t_sub2$Percent.of.population.in.urban.areas)
##
## Coefficients:
##                                   (Intercept)
##                                       -19.0358
## t_sub2$Percent.of.population.in.urban.areas
##                                        0.9002
```

On average, the household penetration rate is -19.04. For every percent increase in the population in urban areas, the household penetration rate increase by 0.90.

Next, we examine the average actual speed.

```
lm(t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Percent.of.population.in.urban.areas)
```

```
##
## Call:
## lm(formula = t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Percent.of.population.in.urban.areas
##
```

```
## Coefficients:
##                                   (Intercept)
##                                      3490.040
## t_sub2$Percent.of.population.in.urban.areas
##                                         5.096
```

On average, the average actual speed is 2451.79. For every percent increase in the population, the average speed increase by 0.005. This is somehow questionable because we expect the more user in the area, the lower the average speed is. The result showed the opposite.

Then we look at the actual average speed with the price for very high speed internet, the high speed,the medium, and the low speed.

```
lm(t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Price.for.very.high.speeds..combined)
```

```
##
## Call:
## lm(formula = t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Price.for.very.high.speeds..combine
##
## Coefficients:
##                                   (Intercept)
##                                       5164.18
## t_sub2$Price.for.very.high.speeds..combined
##                                        -12.31
```

```
lm(t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Price.for.high.speeds..combined)
```

```
##
## Call:
## lm(formula = t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Price.for.high.speeds..combined)
##
## Coefficients:
##                              (Intercept)
##                                 4219.802
## t_sub2$Price.for.high.speeds..combined
##                                   -7.076
```

```
lm(t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Price.for.med.speeds..combined)
```

```
##
## Call:
## lm(formula = t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Price.for.med.speeds..combined)
##
## Coefficients:
##                             (Intercept)
##                                6065.74
## t_sub2$Price.for.med.speeds..combined
##                                  -59.01
```

```
lm(t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Price.for.low.speeds..combined)
```

```
##
## Call:
## lm(formula = t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Price.for.low.speeds..combined)
##
## Coefficients:
##                             (Intercept)
##                                  5126.6
```

```
## t_sub2$Price.for.low.speeds..combined
##                                -46.7
```

On average, the acctual speed for very high speed price is 7526.16. The average speed for high speed price is 5339.31. The average speed for medium speed is 6909.38. This speed is higher than the high speed internet. The average speed for low speed price is 6178.61. Again, the average speed is higher than high speed internet.

Because of the problematic average speed for high price internet we added the percent of population in urban areas.

```
lm(t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Percent.of.population.in.urban.areas + t_sub2$Pri
```

```
##
## Call:
## lm(formula = t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Percent.of.population.in.urban.areas
##     t_sub2$Price.for.very.high.speeds..combined)
##
## Coefficients:
##                                  (Intercept)
##                                    5174.5321
## t_sub2$Percent.of.population.in.urban.areas
##                                      -0.1431
## t_sub2$Price.for.very.high.speeds..combined
##                                     -12.3006
```

The average speed for very high speed internet price is 3788.05. For every dollar less the consumers paid, the speed reduced by 21.84.

```
lm(t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Percent.of.population.in.urban.areas + t_sub2$Pri
```

```
##
## Call:
## lm(formula = t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Percent.of.population.in.urban.areas
##     t_sub2$Price.for.high.speeds..combined)
##
## Coefficients:
##                                  (Intercept)
##                                     3793.016
## t_sub2$Percent.of.population.in.urban.areas
##                                        5.690
##      t_sub2$Price.for.high.speeds..combined
##                                       -7.284
```

The average speed for high speed internet is 3585.35. For every dollar less the consumers paid, the speed reduced by 21.84.

```
lm(t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Percent.of.population.in.urban.areas + t_sub2$Pri
```

```
##
## Call:
## lm(formula = t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Percent.of.population.in.urban.areas
##     t_sub2$Price.for.med.speeds..combined)
##
## Coefficients:
##                                  (Intercept)
##                                     4682.40
## t_sub2$Percent.of.population.in.urban.areas
##                                       21.77
```

```
##       t_sub2$Price.for.med.speeds..combined
##                                     -66.77
```

The average speed for medium speed internet is 4045.76. For every dollar less the consumers paid, the speed reduce by 79.98.

```
lm(t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Percent.of.population.in.urban.areas + t_sub2$Pri
```

```
##
## Call:
## lm(formula = t_sub2$Average.actual.speed..Akamai..kbps. ~ t_sub2$Percent.of.population.in.urban.areas
##     t_sub2$Price.for.low.speeds..combined)
##
## Coefficients:
##                                   (Intercept)
##                                       4210.54
## t_sub2$Percent.of.population.in.urban.areas
##                                         14.93
##       t_sub2$Price.for.low.speeds..combined
##                                     -54.57
```

The average speed for low speed internet is 2987.39. For every dollar less the consumers paid for the internet, the average speed reduce by 83.42.

In conclusion, the average speed is the highest for medium internet speed consumers. The result suggested that the price for high and very high speed did not have the advertised speed the internet providers suggested.

# Conclusion

In this report, we examined speed, household penetration, and price of networks for 30 countries. In the univariate analysis, we found that Korea has extremely good internet. We also found a positive relationship between penetration and speeds; a negative one between speeds and price. It seems that the countries with good internet are not only getting faster speeds, but paying less for it, too. Of course, without accounting for the socio-economic status of these countries, it is hard to draw sweeping conclusions on who is getting ripped off and who is not. I.e. United States prices for very high speeds were highest in this dataset, but this may have contributions from the strength of the dollar.

In addition, we found that countries with higher penetration rate also have higher speed networks; countries with higher speed networks have cheaper prices. There seem to be benefits of open access policies of lowering prices for network owners.

To examine the claim that attempts to regulate prices or mandate greater penetration will reduce incentives for ISP's to invest in non-rural areas, we examined the relationship between price and urban population as a secondary effect. We found a positive relationship: countries with higher population in urban areas showed higher prices. Thus, the open access policies do not seem to hinder any ISP investments. Cheaper networks and higher penetration rates tend to correspond to higher network speeds. We conclude that there is evidence for beneficial effects of open access policies.