An Explanatory Analysis of Broadband Data

Katayoun Borojerdi, Chitra Agastya & Dean Wang

9/21/2018

Introduction

Statement of Research Question

With the rise of the Internet, it has become increasingly important to be able to access online resources quickly. As a result, more and more emphasis has been placed on having better broadband infrastructure and policies that would result in a better high speed Internet market.

Since providing Internet service is a costly venture due to the infrastructure and equipment requirements, there may only be a few entrants to such a market. As a result, Internet service markets may experience phenomena like monopolies or oligopolies if proper regulations are not put in place.

The question we will investigate is whether there is indeed a trade off between Internet penetration, Internet price, and Internet speed across countries. Another question is whether unregulated Internet markets perform better or worse than regulated markets.

Loading Dataset Into R

Our first task is to load the provided `Price.csv`, `Penetration.csv`, and `Speed.csv` files.

```
Penetration = read.csv("Penetration.csv")
Price = read.csv("Price.csv")
Speed = read.csv("Speed.csv")
```

Looking at the various columns, we realized that certain columns that should have been numeric were actually treated as character strings by R. We believe this is because of special characters like "%", "$", and "," included in the data as shown below.

```
head(Penetration$Percent.of.population.in.urban.areas)
```

```
## [1] 88% 66% 97% 80% 74% 86%
## 22 Levels:  162% 56% 58% 59% 61% 66% 67% 68% 74% 75% 76% 77% 80% ... 97%
```

```
head(Price$Price.for.med.speeds..combined)
```

```
## [1]  $61.02   $30.76   $31.61   $40.48   $43.37   $28.82
## 30 Levels:  $23.32   $26.22   $27.91   $28.82   $30.73  ...  $82.76
```

```
head(Speed$Maximum.advertised.speed.OECD..kbps.)
```

```
## [1]  30,000    25,600    20,000    25,000    20,480    100,000
## 16 Levels:   1,000,000   10,240   100,000   110,000   20,000  ...  60,000
```

Further inspecting the data we see that there is also a completely null column in the penetration dataset.

```
head(Penetration$X)
```

```
## [1] NA NA NA NA NA NA
```

Another realization we make is that the Penetration and Speed datasets have 1 and 2 completely empty rows, respectively. Here, we only show this using the first two columns so that we don't take up too much space in the report. You can observe the extra rows by using the tail() function.

```
tail(Penetration[,c(1, 2)])
```

```
##           Country Country.Code
## 26         Sweden           SE
## 27    Switzerland           CH
## 28         Turkey           TR
## 29 United Kingdom           GB
## 30  United States           US
## 31
```

```
tail(Speed[,c(1, 2)])
```

```
##           Country Country.code
## 27    Switzerland           CH
## 28         Turkey           TR
## 29 United Kingdom           GB
## 30  United States           US
## 31
## 32
```

In order to deal with the special characters within numeric data, we use the `read_csv` function from `readr` to automatically clean these special characters to get clean numeric data. We also use this opportunity to skip the completely null column.

First, read the penetration dataset into R using the `read_csv` function from `readr`.

```
library(readr)
Penetration <- read_csv("Penetration.csv",
  col_types = cols(`Growth in 3G penetration` = col_number(),
        `Percent of population in urban areas` = col_number(),
  X13 = col_skip()))
```

To deal with the last row of Penetration being blank, we exclude that row.

```
Penetration = Penetration[-31, ]
```

Next, load the price dataset into R, again using `read.csv` function from `readr` again.

```
Price <- read_csv("Price.csv",
  col_types = cols(`Price for high speeds, combined` = col_number(),
      `Price for low speeds, combined` = col_number(),
      `Price for med speeds, combined` = col_number(),
      `Price for very high speeds, combined` = col_number()))
```

Next, we use the same method to load the speed dataset into R.

```
Speed <- read_csv("Speed.csv")
```

The last two rows of `Speed` are blank, so we exclude these rows.

```
Speed = Speed[-32, ]
Speed = Speed[-31, ]
```

The `Country Code` column in `Speed` is capitalized differently than the other two datasets, so we rename this one to match. This will be useful when we join the data.

```
colnames(Speed)[names(Speed) == "Country code"] <- "Country Code"
```

Join the data together. We use `Country` and `Country Code` since it is a unique identifier and all three datasets have it. We use the `full_join` function available in the `dplyr` library because we would like to keep

all observations from each of the three datasets provided.

```
library(dplyr)
Broadband_Prep = full_join(Penetration,
                           Price, by = NULL, copy = FALSE,
                           suffix = c("Country","Country Code"))

Broadband_Prep = full_join(Broadband_Prep,
                           Speed, by = NULL, copy = FALSE,
                           suffix = c("Country","Country Code"))
```

Dataset description

The fully joined dataset has 31 columns (variables).

```
length(Broadband_Prep)
```

```
## [1] 31
```

The fully joined dataset has 30 rows (observations).

```
nrow(Broadband_Prep)
```

```
## [1] 30
```

Almost all the columns are numeric, with the exception of `Country` and `Country Code` which are character/string data. The list below characterized the list of variables that are available to use to start to analyze and explore.

```
str(Broadband_Prep)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    30 obs. of  31 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 ...
##  $ 2G 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
##  $ 3G 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             : int  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             : int  68 53 42 62 39 43 48 89 61 68 ...
##  $ 90p. Download, speedtest.net (kbps)       : num  10676 12338 14911 9356 11493 ...
##  $ 90p. Upload, speedtest.net (kbps)         : num  853 1467 922 960 3912 ...
##  $ 10p. Latency, speedtest.net               : int  22 18 18 19 10 11 14 34 21 26 ...
```

Data Quality

We took a summary of the data, and saw that overall most variables are fully populated, with just four columns having missing values. The columns with missing data are `Price for low speeds, combined`, `Price for high speeds, combined`,`Price for very high speeds, combined` and `Average actual speed, akamai (kbps)` which we will further investigate in the univariate analysis. It is important to note that Columns that were strings before because of special characters are now numeric. Below we show an example summary of one variable as to not take up too much space in this analysis but we did do a preliminary summary of all variables before starting our analysis in order to find any obvious anomalies which we will discuss in their respective sections.

```
summary(Broadband_Prep$`Penetration per 100 OECD, 2008`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.20   17.68   25.60   23.96   30.51   37.18
```

Univariate Analysis of Key Variables

In this section we will perform individual exploratory analysis of each of the three variables of interest, speed, penetration and price. We first look at Speed and Penetration because they are the two outcome variables which metrics on performance of a network industry success are measured on.

Speed

The speed dataset provided has several variables including advertised speeds (avg and max.), actual speeds (avg only) as well as speedtest.net data. The speedtest.net data is comprised of download, upload and latency averages, medians and standard deviations for each country.

Note: We will not be using the 90p. Download, 90p Upload and 10p. and Latency datasets in our analysis. We did a preliminary explanatory analysis of these variables with findings that they are highly correlated with the median and average speed data for upload, downloand and latency data. As a result, in addition to research on speedtest.net we found that the 90p, 10p stands for 90 percentile and 10 percentile of measured test data, respectively, and that by using Median (50 percentile) or Average would serve as a representative sample for this analysis going forward.

Next, to get started we quickly investigate how many missing values the speed variables have using our Pre-joined speed dataset.

```
sum(is.na(Speed))
```

```
## [1] 4
```

Since there are 4 missing values we need to investigate each variable to find out which one(s) has missing values. We do this by going down the list of speed variables, one by one to see if we can see which one has the missing values.

```
sum(is.na(Broadband_Prep$`Maximum advertised speed OECD (kbps)`))
```

```
## [1] 0
```

```
sum(is.na(Broadband_Prep$`Average advertised speed OECD (kbps)`))
```

```
## [1] 0
```

```
sum(is.na(Broadband_Prep$`Average actual speed, Akamai (kbps)`))
```

```
## [1] 4
```

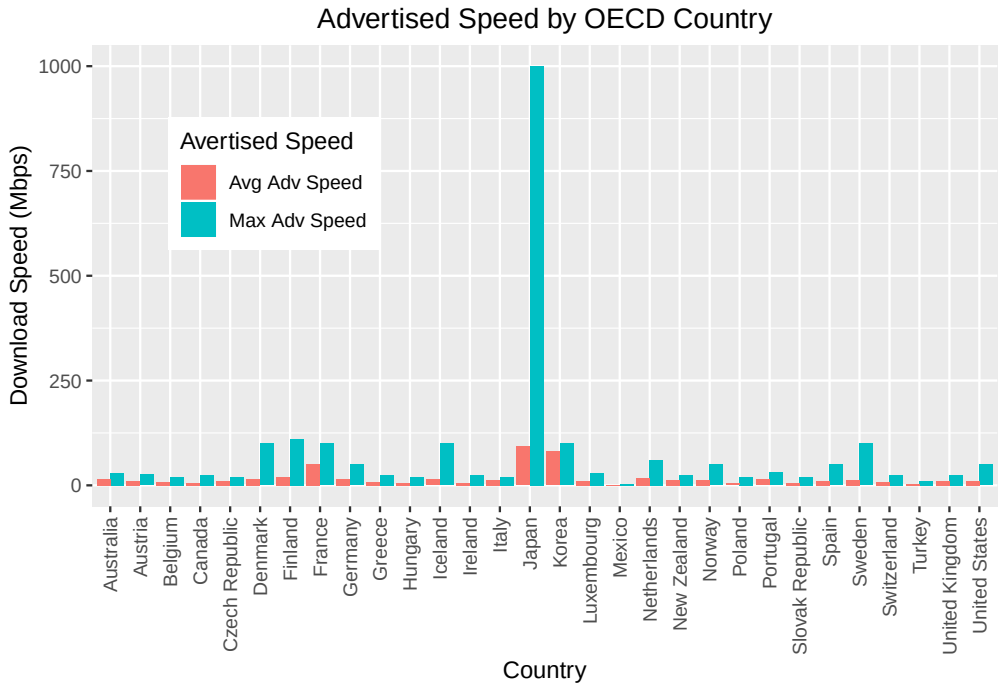
Since we see that the number of NA values in the Average Actual Speed dataset is 4 it confirms that this is the variable with missing values and something to consider when using this variable which may not allow us to maximize the number of observations for our upcoming analysis.

Now that we have evaluated the completeness of data we can move on to investigate the data quality and look for any outliers that may exist in each of the speed variables.

Starting with maximum and average advertised speed variables we create a bar graph to visualize the data.

```
library(ggplot2)
country = rep(c(Broadband_Prep$Country),2)
adv_speed = c(rep("Avg Adv Speed", 30), rep("Max Adv Speed", 30))
value = c(Broadband_Prep$`Average advertised speed OECD (kbps)`/1000,
          Broadband_Prep$`Maximum advertised speed OECD (kbps)`/1000)
data_plot = data.frame(country, adv_speed, value)

ggplot(data_plot, aes(fill=adv_speed, y=value, x=country)) +
  geom_col(position="dodge") + theme(axis.text.x = element_text(angle=90, hjust = 1, vjust = 0.5)) +
  labs(y="Download Speed (Mbps)", x = "Country") + ggtitle("Advertised Speed by OECD Country") +
  theme(plot.title = element_text(hjust = 0.5)) + guides(fill=guide_legend(title="Avertised Speed")) +
  theme(legend.position = c(0.2, 0.7))
```



One obvious observation is the max advertised speed in Japan is more than 4 times any other data point. We

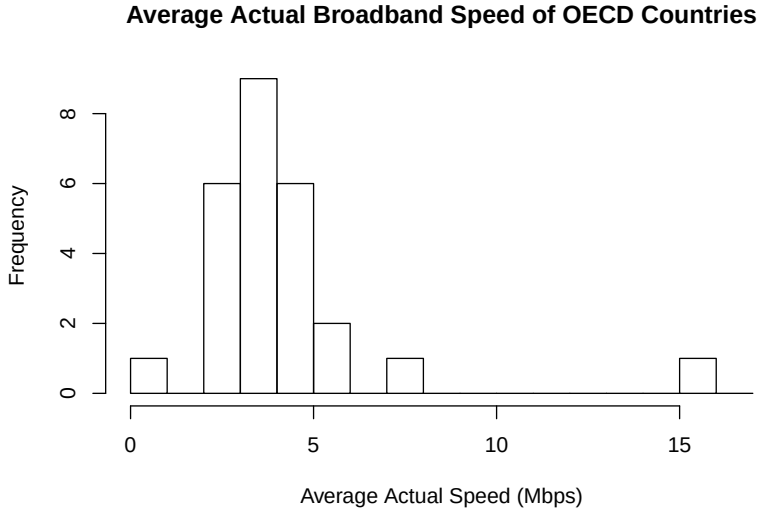
believe this is unlikely to be true representation of Japan's advertised speed and we decide this may be one outlive to ignore.

Note, after some research we did find out that around 2008 there was one company in Japan advertising ~1Gbps of download, however unrealistic that appears for the time.

After considering average and max advertised speeds we came to the conclusion that we should not use these variables as our speed indicator variable because it does not adequately represent what customers obtain for speed. Advertised speeds, regardless of average or max are consistently much higher than actual speeds measured (see Analysis of Secondary affects : Advertised vs. Actual Speed). For this reason we believe one of the speed variables from the speedtest.net source is probably a better representation of speed.

Before moving on to the speedtest.net data there is one more speed variable to consider. We can look at a histogram of average actual speed.

```
hist(Broadband_Prep$`Average actual speed, Akamai (kbps)`/1000, breaks = c(0:17),
     main="Average Actual Broadband Speed of OECD Countries", xlab="Average Actual Speed (Mbps)")
```




```
summary (Broadband_Prep$`Average actual speed, Akamai (kbps)`)
```

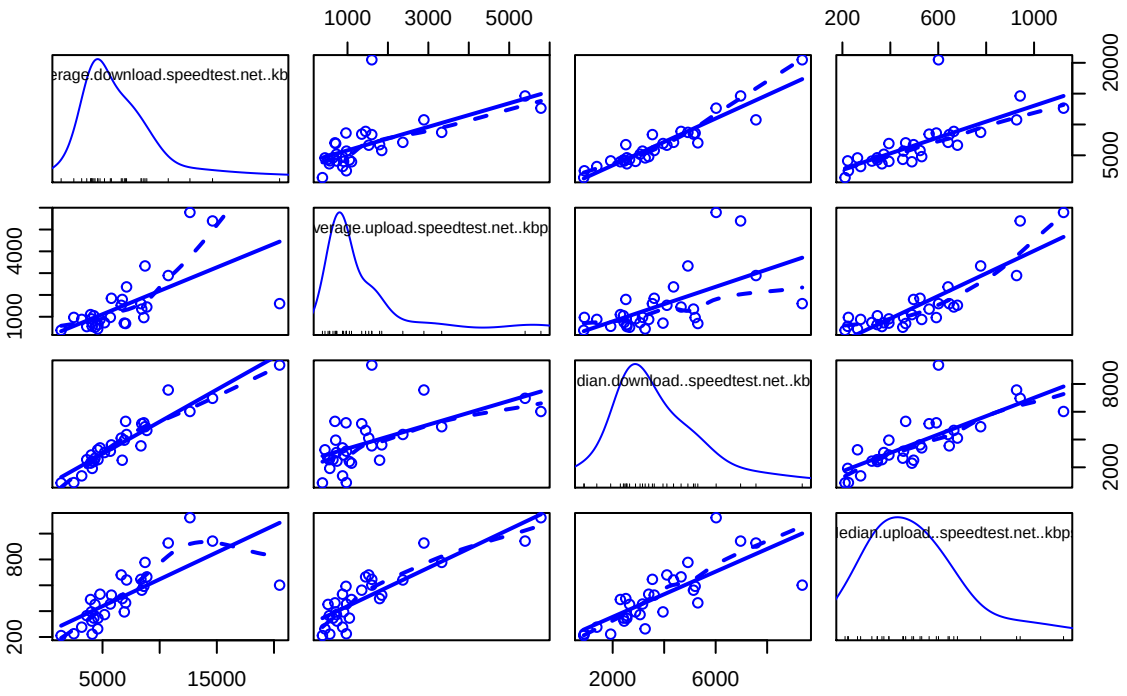
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     948    3032    3780    4205    4474   15239       4
```

The histogram of average actual speed data indicates a unimodal, slightly right skew distribution. There are a few outliers near the min and max that we would investigate further if we wanted to use this variable but because we found out earlier that this speed variable has four missing values and we do not want to reduce our number of observations by four, we are going to investigate other variables.

Now we start to explore the speedtest.net data. first, because network speed is typically evaluated by upload and download speed, we will look at these variables to determine their relationships to each other as well as whether the correlation that we would expect is present.

```
library(car)
scatterplotMatrix(~`Average download speedtest.net (kbps)` +
                    `Average upload speedtest.net (kbps)` +
                    `Median download, speedtest.net (kbps)` +
                    `Median upload, speedtest.net (kbps)`,
                  data= Broadband_Prep, main="Average and Median, Download and Upload Speeds",
                  smooth=list(spread=FALSE))
```

Average and Median, Download and Upload Speeds



From this scatterplot matrix a list of observations can be made:

1. For all four variables, average and median, download and upload speed, have a strong positive correlation.

2. The strongest correlation appears to be between average and median download speed variables.
3. The distribution of data in there variables is all unimodal and with right skew.

Note: it is important to point out the axis scale. While the correlations are strong between these variables, the upload and download speeds are on different scales in terms of magnitude, with download speeds being much higher.

While we believe the download speed data is most representative candidate of overall speed, we will check the correlations between the four upload and download variables.

```
cor (Broadband_Prep$`Average download speedtest.net (kbps)`,
     Broadband_Prep$`Median download, speedtest.net (kbps)`)
```

```
## [1] 0.937447
```

```
cor (Broadband_Prep$`Average upload speedtest.net (kbps)` ,
     Broadband_Prep$`Median upload, speedtest.net (kbps)`)
```

```
## [1] 0.8775169
```

```
cor (Broadband_Prep$`Average download speedtest.net (kbps)`,
     Broadband_Prep$`Average upload speedtest.net (kbps)`)
```

```
## [1] 0.6403798
```

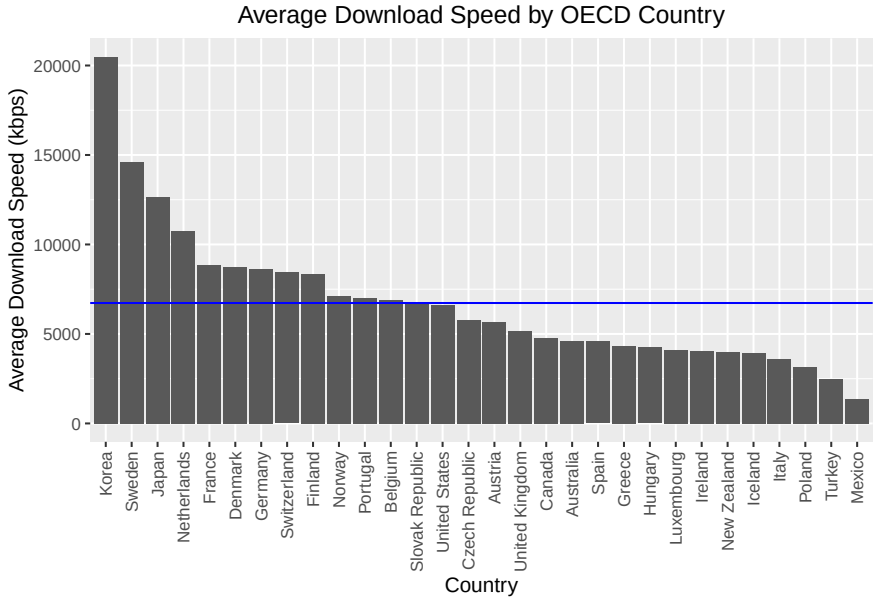
```
cor (Broadband_Prep$`Median download, speedtest.net (kbps)`,
     Broadband_Prep$`Median upload, speedtest.net (kbps)`)
```

```
## [1] 0.7612491
```

Now, we can confirm our observation from the scatterplot matrix that average download and median variables have the strongest correlation at 0.93. Next we note that average and median upload speeds also have a strong correlation at 0.88. Based on these high correlations and in addition to the fact that most users use their network connection for downloading we can operationalize the concept of speed using 'Average download speedtest.net (kbps)' for future analysis, keeping in mind it is representative of upload speed as well.

With our speed variable decided, we can look at how the average download speed compares between the various OECD countries.

```
ggplot(data = Broadband_Prep,
       aes(x=reorder(Broadband_Prep$Country,
                     -Broadband_Prep$`Average download speedtest.net (kbps)`),
           y= `Average download speedtest.net (kbps)`)) +
  geom_bar(stat="identity")  +
  geom_hline(yintercept = mean(Broadband_Prep$`Average download speedtest.net (kbps)`),
             color="blue") +
  theme(axis.text.x = element_text(angle=90, hjust = 1, vjust = 0.5)) +
  labs(y="Average Download Speed (kbps)", x="Country") +
  ggtitle("Average Download Speed by OECD Country") +
  theme(plot.title = element_text(hjust = 0.5))
```

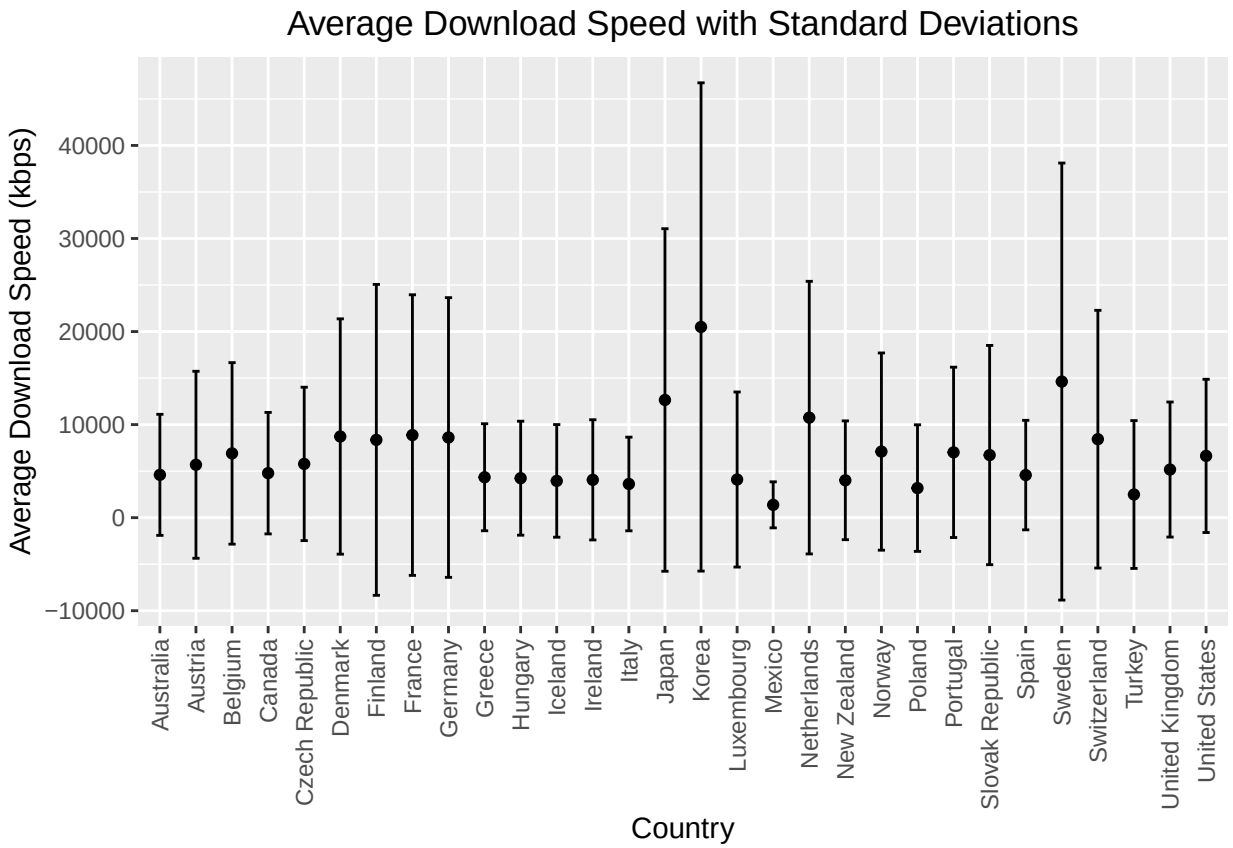


Based on the bar graphs above: 1. Almost 60% of the countries have less than the average download speed of 7Mbps 2. Countries like United States, Slovak Republic and Mexico, that have not opted for open access regulation have download speeds less than the average download speeds. 3. Korea, Sweden and Japan lead the OECD countries in fastest average download speeds, distinguishing themselves from the pack.

Lastly, as stated earlier we know that average download speed data is important; this is because most users are using the internet for downloading content, for example by web browsing or streaming data. We wanted to explore one more aspect of the average download data, but this time including the standard deviation to see if the trends with these countries stays consistent.

```
library(ggplot2)
ggplot(Broadband_Prep, aes(x=`Country`, y=`Average download speedtest.net (kbps)`)) +
     geom_errorbar(aes(ymin=`Average download speedtest.net (kbps)` -
                           `Standard deviation download, speedtest.net`,
                       ymax=`Average download speedtest.net (kbps)` +
                           `Standard deviation download, speedtest.net`), width=.2) +
```

```
    geom_line() + geom_point() + theme(axis.text.x = element_text(angle=90, hjust = 1, vjust = 0.5)) +
    labs(y="Average Download Speed (kbps)", x="Country") +
    ggtitle("Average Download Speed with Standard Deviations") +
    theme(plot.title = element_text(hjust = 0.5))
```



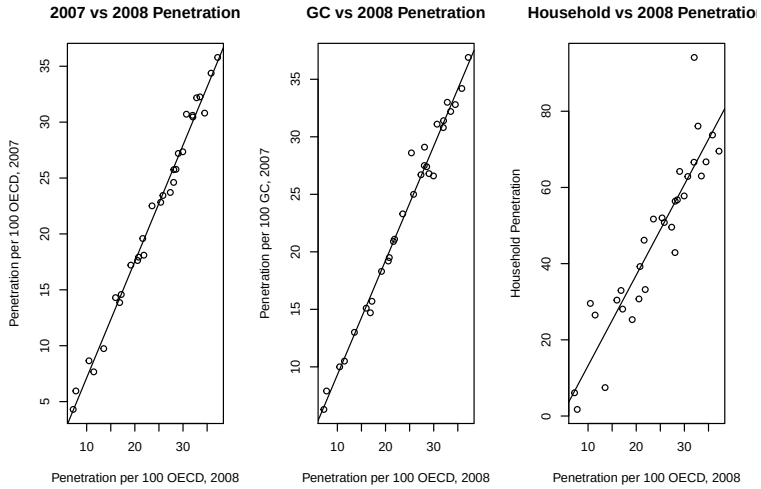
Based on the graph shown above we can see that countries such as South Korea, Sweden and Japan do offer the highest download speeds but they also have the broadest range of download speeds in terms of standard deviation. Again, as for the countries who have not adopted any form of open access such as the United States, Slovak Republic and Mexico, the first two are averaging within the middle of the pack of countries while Mexico shows the lowest range of average download speed.

Penetration

We have several variables in the penetration dataset to investigate and choose between to determine which variable to use to operationalize the penetration concept. Initially we suspect that there may be a relationship between four of the overall penetration values: penetration per 100 in 2008, penetration per 100 in 2007, and penetration per 100 GC and Household Penetration variables. We have a preference to use the latest dataset from 2008 so for this reason we plot the other three against the penetration per 2008 variable to explore correlations.

```
attach (mtcars)
par(mfrow=c(1,3))
plot(Broadband_Prep$`Penetration per 100 OECD, 2008`,
     Broadband_Prep$`Penetration per 100 OECD, 2007`,
     main="2007 vs 2008 Penetration",
     ylab="Penetration per 100 OECD, 2007",
     xlab="Penetration per 100 OECD, 2008")
```

```
abline(lm(Broadband_Prep$`Penetration per 100 OECD, 2007` ~
            Broadband_Prep$`Penetration per 100 OECD, 2008`))
plot(Broadband_Prep$`Penetration per 100 OECD, 2008`,
     Broadband_Prep$`Penetration per 100 GC`,
     main="GC vs 2008 Penetration",
     ylab="Penetration per 100 GC, 2007",
     xlab="Penetration per 100 OECD, 2008")
abline(lm(Broadband_Prep$`Penetration per 100 GC` ~
            Broadband_Prep$`Penetration per 100 OECD, 2008`))
plot(Broadband_Prep$`Penetration per 100 OECD, 2008`,
     Broadband_Prep$`Household penetration, OECD`,
     main="Household vs 2008 Penetration",
     ylab="Household Penetration", xlab="Penetration per 100 OECD, 2008")
abline(lm(Broadband_Prep$`Household penetration, OECD` ~
            Broadband_Prep$`Penetration per 100 OECD, 2008`))
```




```
cor(Broadband_Prep$`Penetration per 100 OECD, 2007`,
    Broadband_Prep$`Penetration per 100 OECD, 2008`)
```

```
## [1] 0.9945567
```

```
cor(Broadband_Prep$`Penetration per 100 GC`,
    Broadband_Prep$`Penetration per 100 OECD, 2008`)
```

```
## [1] 0.9912493
```

```
cor(Broadband_Prep$`Household penetration, OECD`,
    Broadband_Prep$`Penetration per 100 OECD, 2008`)
```

```
## [1] 0.920708
```

As we show above, the three are very highly correlated with each other with a value of 0.92 at the lowest, which indicates that we may use penetration rate for 2008 as our variable for further analysis.

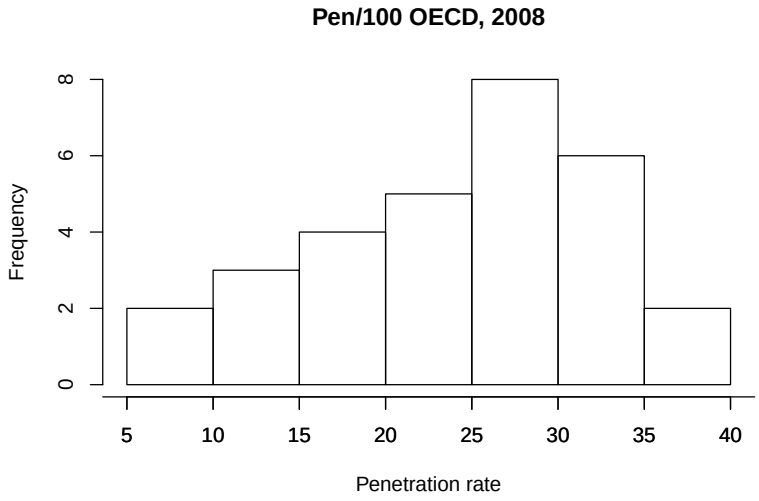
We now look at the summary and distribution of the penetration rate for 2008 as well as visualize it by OECD countries to explore any anomalies or relationships.

```
pen=Broadband_Prep$`Penetration per 100 OECD, 2008`
summary(pen)
```

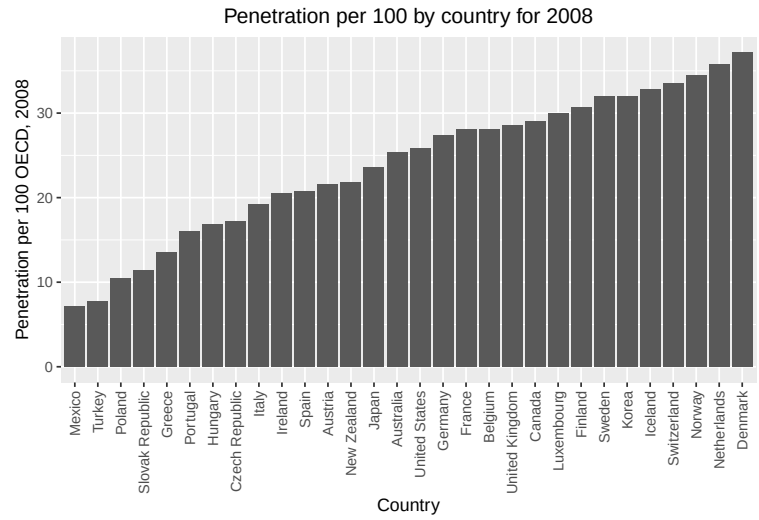
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    7.20   17.68   25.60   23.96   30.51   37.18
```

```
hist(pen,
     main="Pen/100 OECD, 2008",
     xlab="Penetration rate")
axis(side=1, at=seq(0,100,5))
```




```
library(ggplot2)
ggplot(data = Broadband_Prep, aes(x = reorder(Broadband_Prep$Country,
                                              Broadband_Prep$`Penetration per 100 OECD, 2008`),
                                  y= Broadband_Prep$`Penetration per 100 OECD, 2008`)) +
  geom_bar(stat="identity") + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
  labs(y="Penetration per 100 OECD, 2008", x="Country") +
  ggtitle("Penetration per 100 by country for 2008") + theme(plot.title = element_text(hjust = 0.5))
```



The histogram for 'Penetration rate per 100 OECD, 2008' shown above seems to have a left skew. We can transform this using a square transformation to perform any future regression analysis.

Here are the observations we make:

1. Only up to 37 per 100 people in the OECD countries have access to internet
2. The country with the lowest penetration rate, Mexico, has only 7 out of every 100 people that have access to internet. This country also happens to not have open access regulations

3. The average penetration rate is at ~23%. Since median is at 25.6% and is larger than the mean, it indicates that more than half of our observations have a penetration rate larger than 23%

4. The bar graph above shows countries ordered by penetration rate. We see that again, Mexico, and Slovak Republic, both of who have not adopted open access regulations are among the countries with lowest penetration. The US also is just around the median value of penetration rate

Next we see if we can gain any insights from growth in penetration in 2007 vs 2008

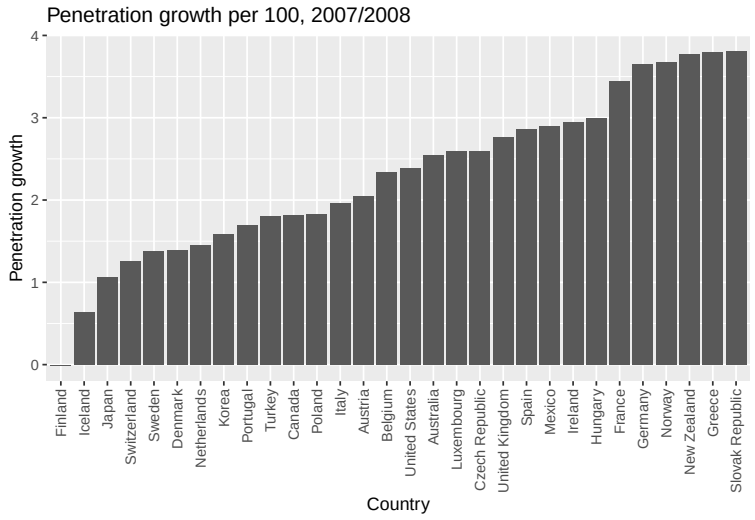
```
growth = (Broadband_Prep$`Penetration per 100 OECD, 2008` -
            Broadband_Prep$`Penetration per 100 OECD, 2007`)
summary(growth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -0.010   1.607   2.365   2.300   2.938   3.810
```

```
growth
```

```
##  [1]  2.54  2.05  2.34  1.82  2.60  1.39 -0.01  3.44  3.65  3.80  2.99
## [12]  0.64  2.95  1.96  1.06  1.58  2.60  2.90  1.45  3.77  3.68  1.83
## [23]  1.69  3.81  2.86  1.38  1.26  1.81  2.76  2.39
```

```
library(ggplot2)
ggplot(data = Broadband_Prep, aes(x = reorder(Broadband_Prep$Country, growth), y= growth)) +
  geom_bar(stat="identity") + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
  labs(y="Penetration growth", x="Country") + ggtitle("Penetration growth per 100, 2007/2008")
```



From the graph above it looks like countries like Mexico and Slovak Republic have witnessed among higher growth compared to some other countries. But so have many countries with open access regulation. The data is not sufficient to comment on what factors might have caused the growth.

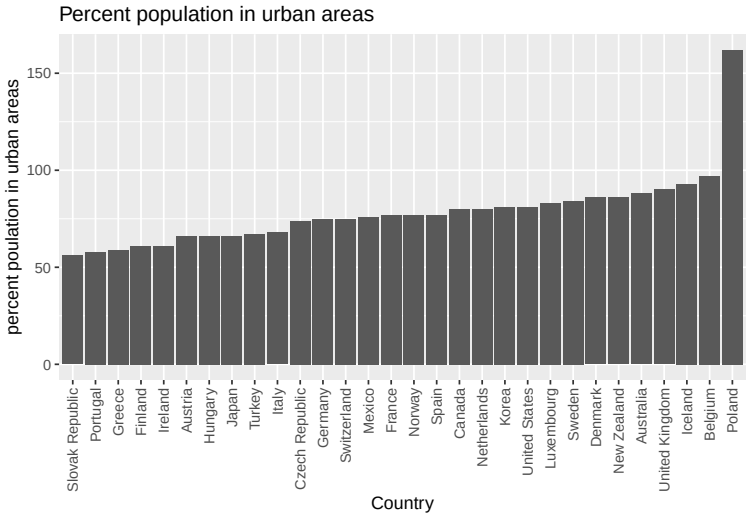
We next look at penetration by urban population

```
growth = (Broadband_Prep$`Penetration per 100 OECD, 2008` -
            Broadband_Prep$`Penetration per 100 OECD, 2007`)
library(ggplot2)
Broadband_Prep$`Percent of population in urban areas`
```

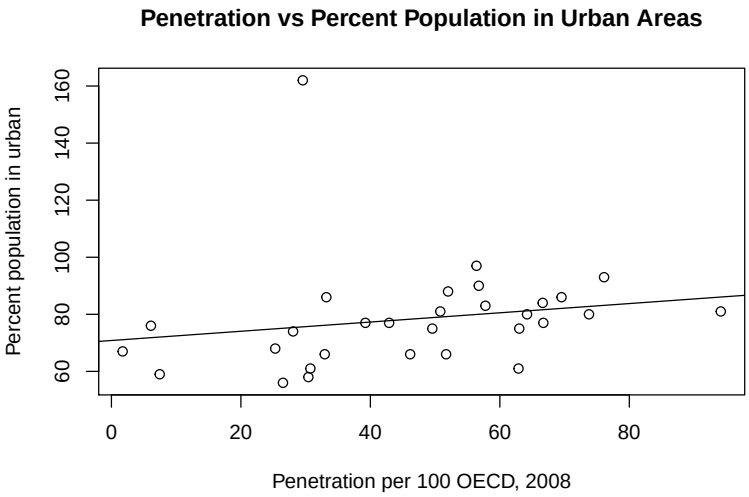
```
##  [1]  88  66  97  80  74  86  61  77  75  59  66  93  61  68  66  81  83
## [18]  76  80  86  77 162  58  56  77  84  75  67  90  81
```

```
ggplot(data = Broadband_Prep,
       aes(x = reorder(Broadband_Prep$Country,
```

```
                        Broadband_Prep$`Percent of population in urban areas`),
                                y= Broadband_Prep$`Percent of population in urban areas`)) +
  geom_bar(stat="identity") + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5)) +
  labs(y="percent poulation in urban areas", x="Country") + ggtitle("Percent population in urban areas")
```




```
plot(Broadband_Prep$`Household penetration, OECD`,
     Broadband_Prep$`Percent of population in urban areas`,
     main="Penetration vs Percent Population in Urban Areas",
     ylab="Percent population in urban", xlab="Penetration per 100 OECD, 2008")
abline(lm(Broadband_Prep$`Percent of population in urban areas` ~ Broadband_Prep$`Household penetration
```




```
cor(Broadband_Prep$`Percent of population in urban areas`,
    Broadband_Prep$`Household penetration, OECD`)
```

```
## [1] 0.1862987
```

There is a very small positive correlation between household penetration and percent population in urban areas.

We compared different penetration variables with the percent of population in urban areas to see if there are any strong correlations

```
urban=Broadband_Prep$`Percent of population in urban areas`
cor(Broadband_Prep$`2G and 3G penetration per 100, OECD`, urban )
```

```
## [1] -0.1405324
```

```
cor(Broadband_Prep$`3G penetration per 100`, urban)
```

```
## [1] -0.07872161
```

```
cor(Broadband_Prep$`Wi-Fi hotspots per 100,000, JiWire`, urban)
```

```
## [1] -0.01676172
```

We don't see any significant correlation between penetration across different technologies and percentage of population in urban areas.

Price

Broadband price is a subset of the broadband data that we explored. Quickly we observed that price data was segmented by four different variables that are characterized by speed being offered. One caveat to note is speeds are identified as low, medium, high and very high speeds but there was no information regarding what exact speeds represent each category. Rather we assume these terms are relative. We also do not know whether the definition of low, medium, high and very high speed are consistent across all countries.

Before attempting any visualization we used the code below to survey any missing values in the price variables:

```
sum(is.na(Broadband_Prep$`Price for low speeds, combined`))
```

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

```
sum(is.na(Broadband_Prep$`Price for med speeds, combined`))
```

```
## [1] 0
```

```
sum(is.na(Broadband_Prep$`Price for high speeds, combined`))
```

```
## [1] 2
```

```
sum(is.na(Broadband_Prep$`Price for very high speeds, combined`))
```

```
## [1] 11
```

Our findings indicate that there is only one NA for low speed, none for medium speed, two for high speed and eleven missing values for the very high speed price variable. Discovering that 11 of the 30 observations are missing data points, approximately ~30% of the data, lets us know that using the very high speed price variable may not be a good representation of prices in this category. However it could also be that very high speed broadband data infrastructure is not available in these countries.

Next based on the above finding we chose to summarize the price variables to help select one as our key price variable to continue exploring.

```
summary(Broadband_Prep$`Price for low speeds, combined`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   13.10   24.01   27.28   29.11   31.96   60.23       1
```

```
summary(Broadband_Prep$`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
```

```
summary(Broadband_Prep$`Price for high speeds, combined`)
```

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

```
summary(Broadband_Prep$`Price for very high speeds, combined`)
```

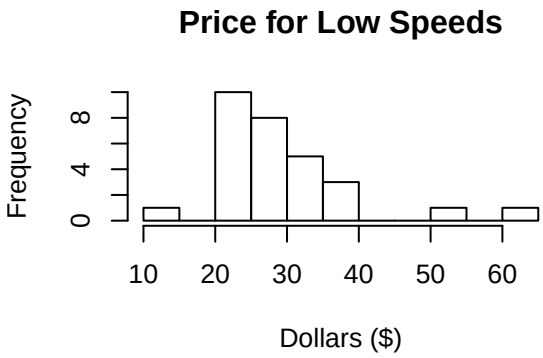
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   32.61   48.04   76.22   77.07  102.92  130.21      11
```

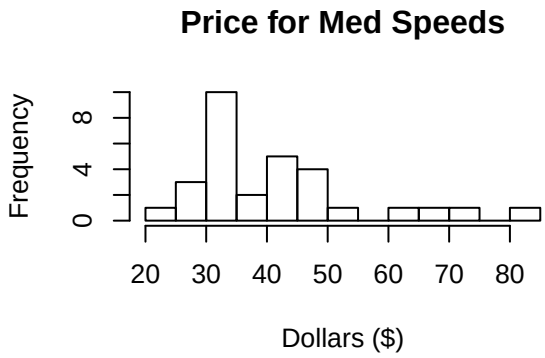
Quickly we can see that the minimum and maximum prices charged overall are going up linearly based on increasing broadband speeds except for the high speed variable, which is $0.69 min and $210.36 max. We note that these data points are suspect and may be driven by an incorrect value and we will need to explore further if we choose to continue to use it in our analysis.

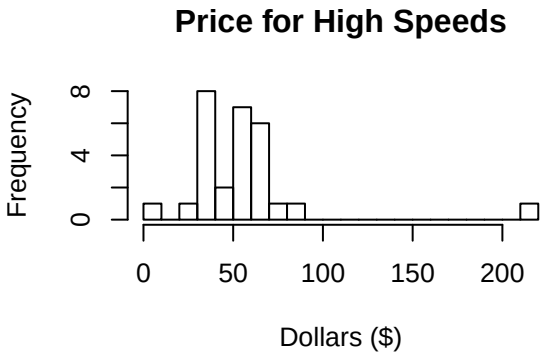
Next we do a check of comparing Median to Mean for each variable in order to assess any indications of issues with the distribution of the data points. Most look fairly close making a note that the price for medium speed has the highest difference indicating possibly the largest variation in distribution.

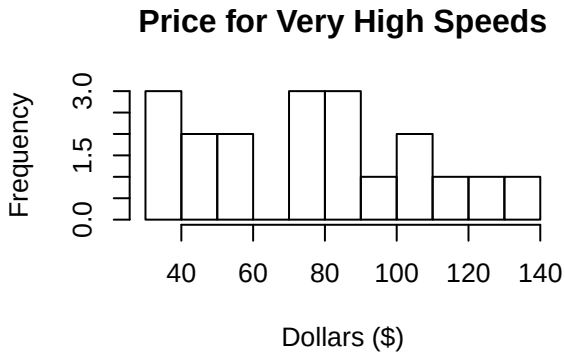
In order to further inspect the distribution of the price variables we have decided to create a histogram of each variable. The following code create a histogram, while trying to keep the x axis as consistent as possible where breaks are between $5 to $10 each so that the plots can be easily compared.

```
par(mfrow = c(2,2))
hist(Broadband_Prep$`Price for low speeds, combined`, breaks = 10,
     main="Price for Low Speeds", xlab="Dollars ($)")
hist(Broadband_Prep$`Price for med speeds, combined`, breaks = 18,
     main="Price for Med Speeds", xlab="Dollars ($)")
hist(Broadband_Prep$`Price for high speeds, combined`, breaks = 20,
     main="Price for High Speeds", xlab="Dollars ($)")
hist(Broadband_Prep$`Price for very high speeds, combined`, breaks = 10,
     main="Price for Very High Speeds", xlab="Dollars ($)")
```









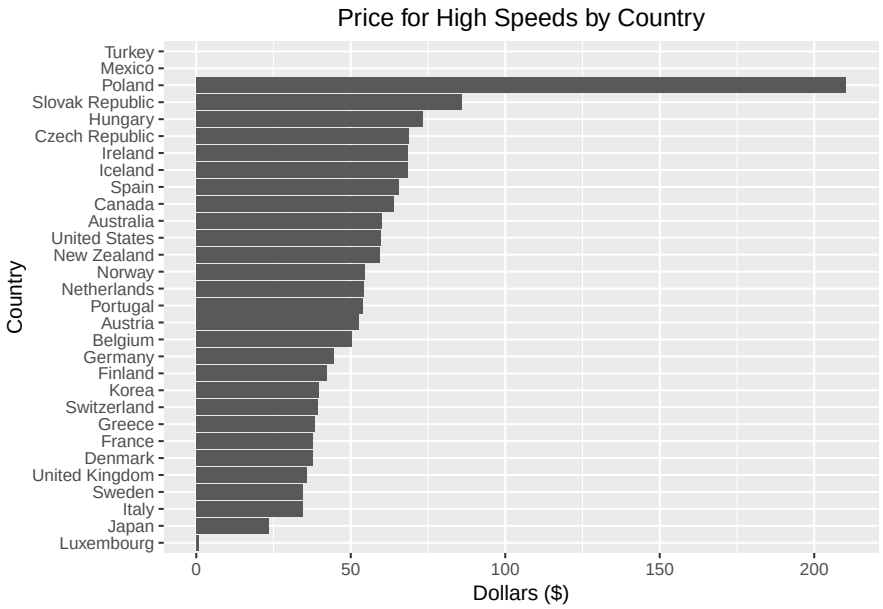
Findings from the basic histogram shown above indicate a few points about each price variable:

1. With the exception of a few outliers people pay somewhere between $20 to $40 for low speed connections.
2. There is more variability in the data of price for medium speed data compared to low but on average one third of countries pay between $30 to $35.
3. In the price for high speed variable we can now confirm that min and max outliers that we suspected earlier when looking at the variable summary. Keeping in mind to remove these outliers this is probably the best price dataset to use in subsequent analysis for price. The distribution is tight and closest to a normal distribution.
4. The Price for very high speeds variable has a wide distribution which shows that prices for very high speeds are tremendously variable and highly depending on the country you live in. We believe this is where private competition and investment in infrastructure make the difference.

Next, we will explore the price for high speed variable across the OECD countries for any findings.

```
library(ggplot2)
# Basic barplot
p <- ggplot(data = Broadband_Prep, aes(x = reorder(Price$Country,
                                                   Broadband_Prep$`Price for high speeds, combined`),
                                       y= `Price for high speeds, combined` )) +
  geom_bar(stat="identity") + labs(y="Dollars ($)", x="Country") + ggtitle("Price for High Speeds by Co
  theme(plot.title = element_text(hjust = 0.5))

# Horizontal bar plot
p + coord_flip()
```



Analysis of Key Relationships

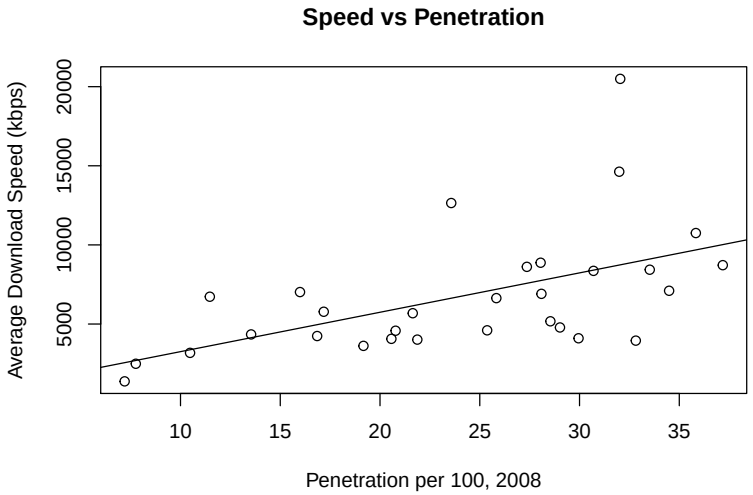
Now that we have thoroughly explored each type of data provided and we have methodically investigated anomalies and selected our best variable to operationalize the speed, penetration and price variables of broadband data we will further investigate the bivariate relationship between these variables.

Speed vs Penetration

We begin our bivariate analysis by comparing key outcome variables: speed vs penetration. These variables are often used to establish whether a given broadband market is a "success" or not. If higher speeds and higher penetrations are found in a market, this can be considered more desirable than the opposites. We investigate whether these variables are related to each other.

We plot the speed variable we have chosen against the penetration one. We see that there is positive correlation between the speed and penetration.

```
plot(Broadband_Prep$`Penetration per 100 OECD, 2008`,
     Broadband_Prep$`Average download speedtest.net (kbps)`,
     main="Speed vs Penetration", ylab="Average Download Speed (kbps)",
     xlab="Penetration per 100, 2008")
abline(lm(Broadband_Prep$`Average download speedtest.net (kbps)` ~
            Broadband_Prep$`Penetration per 100 OECD, 2008`))
```



To get a numerical measure of this, we find the correlation. The correlation agrees with our visual analysis, with a high positive correlation of 0.54 being found.

```
cor(Broadband_Prep$`Penetration per 100 OECD, 2008`,
    Broadband_Prep$`Average download speedtest.net (kbps)`)
```

```
## [1] 0.5398292
```

To summarize our finding, we have found that there is a high positive correlation between speed and penetration. In countries with high penetration, there is higher speed.

There may be several different reasons for this: - higher penetration might cause there to be more infrastructure set up to support the greater online population, leading to better speeds - higher speeds might make getting online more attractive, increasing penetration - other factors like government policies might increase spending in Internet infrastructure, increasing both speeds and penetration

Follow up statistical analysis can be done to test these and other hypothesis of our finding.

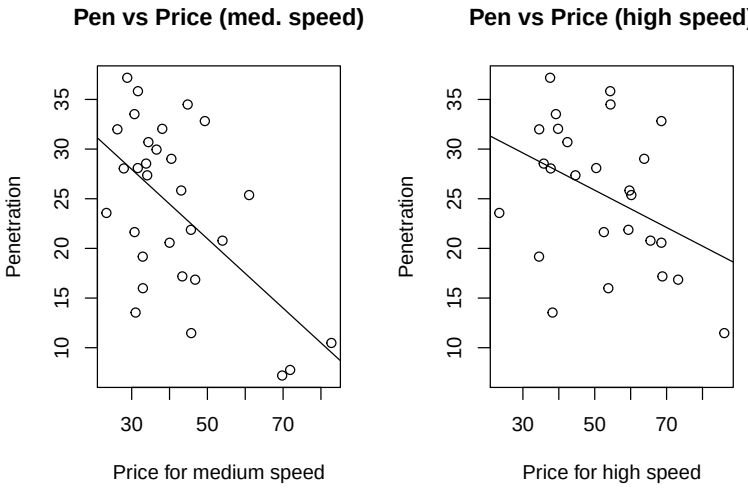
Price and Penetration

We look at how our outcome variable network penetration is affected by broadband price. Since penetration could be different at different price points, we look at relationships between penetration and price for both medium and high speed networks. Once we adjust the price for high speed for the two outliers at the two extremes (as mentioned in earlier sections), it makes the price ranges for medium and high speeds comparable ($20-$90). Since many countries do not have very high speed network, we will not consider the price variable for very high speed networks.

We begin by examining the correlation between these two variables. The scatter plot shows us that there is an overall negative linear relationship between penetration and price at medium and high speeds.

```
newHigh = Broadband_Prep$`Price for high speeds, combined`
newHigh = newHigh[c(-17,-22)]
par(mfrow = c(1,2))
plot(Broadband_Prep$`Price for med speeds, combined`,
     Broadband_Prep$`Penetration per 100 OECD, 2008`,
     xlab = "Price for medium speed", ylab = "Penetration",
     main = "Pen vs Price (med. speed)")
abline(lm(Broadband_Prep$`Penetration per 100 OECD, 2008` ~
            Broadband_Prep$`Price for med speeds, combined`))
plot(newHigh,
```

```
     Broadband_Prep$`Penetration per 100 OECD, 2008`[c(-17,-22)],
     xlab = "Price for high speed", ylab = "Penetration",
     main = "Pen vs Price (high speed)")
abline(lm(Broadband_Prep$`Penetration per 100 OECD, 2008`[c(-17,-22)] ~
            newHigh))
```




```
cor(Broadband_Prep$`Price for med speeds, combined`,
  Broadband_Prep$`Penetration per 100 OECD, 2008`,
  use="complete.obs")
```

```
## [1] -0.5842712
```

```
cor(newHigh,
  Broadband_Prep$`Penetration per 100 OECD, 2008`[c(-17,-22)],
  use="complete.obs")
```

```
## [1] -0.3952233
```

Also we can see that the correlation is negative between penetration and price for both medium and high speed networks. This indicates the lower the price, the higher the penetration.

```
print("Medium price summary: ")
```

```
## [1] "Medium price summary: "
```

```
summary(Broadband_Prep$`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
```

```
print("High price summary: ")
```

```
## [1] "High price summary: "
```

```
summary(newHigh)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   23.43   38.48   53.16   51.81   62.88   86.06       2
```

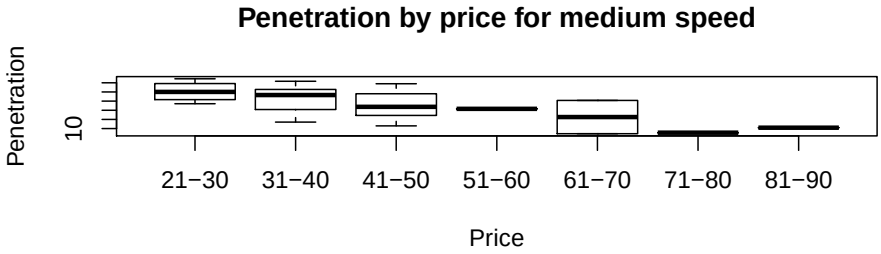
The minimum and maximum prices for medium speed are $23 and $83 respectively. So we divide price into intervals of 10 starting from $20 to $90 to see how the penetration looks across these different price intervals. Similarly for prices for high speed, we divide the price into intervals of $10 starting from $20 to $90 to cover the entire distribution. We then use a box plot to see the penetration in these price bins.

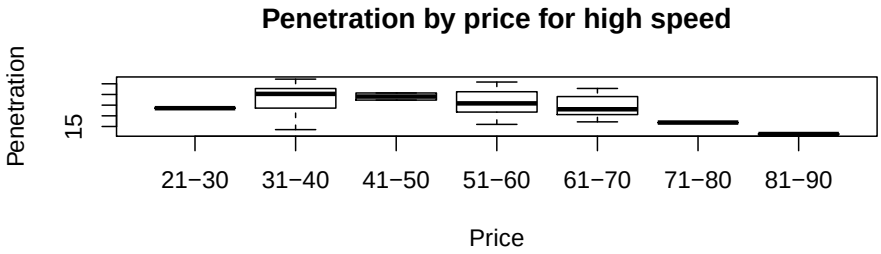
```
priceMedium = cut(Broadband_Prep$`Price for med speeds, combined`,
                  breaks = seq(20,90,10),
                  labels = c("21-30", "31-40",
                             "41-50","51-60","61-70", "71-80", "81-90"))

priceHigh1 = cut(Broadband_Prep$`Price for high speeds, combined`,
                  breaks = seq(20,90,10),
                  labels =  c("21-30", "31-40",
                             "41-50","51-60","61-70", "71-80", "81-90"))

par(mfrow = c(2,1))
boxplot(Broadband_Prep$`Penetration per 100 OECD, 2008` ~ priceMedium,
        data = Broadband_Prep,
        main = "Penetration by price for medium speed",
        xlab = "Price", ylab = "Penetration")

boxplot(Broadband_Prep$`Penetration per 100 OECD, 2008` ~ priceHigh1,
        data = Broadband_Prep,
        main = "Penetration by price for high speed",
        xlab = "Price", ylab = "Penetration")
```





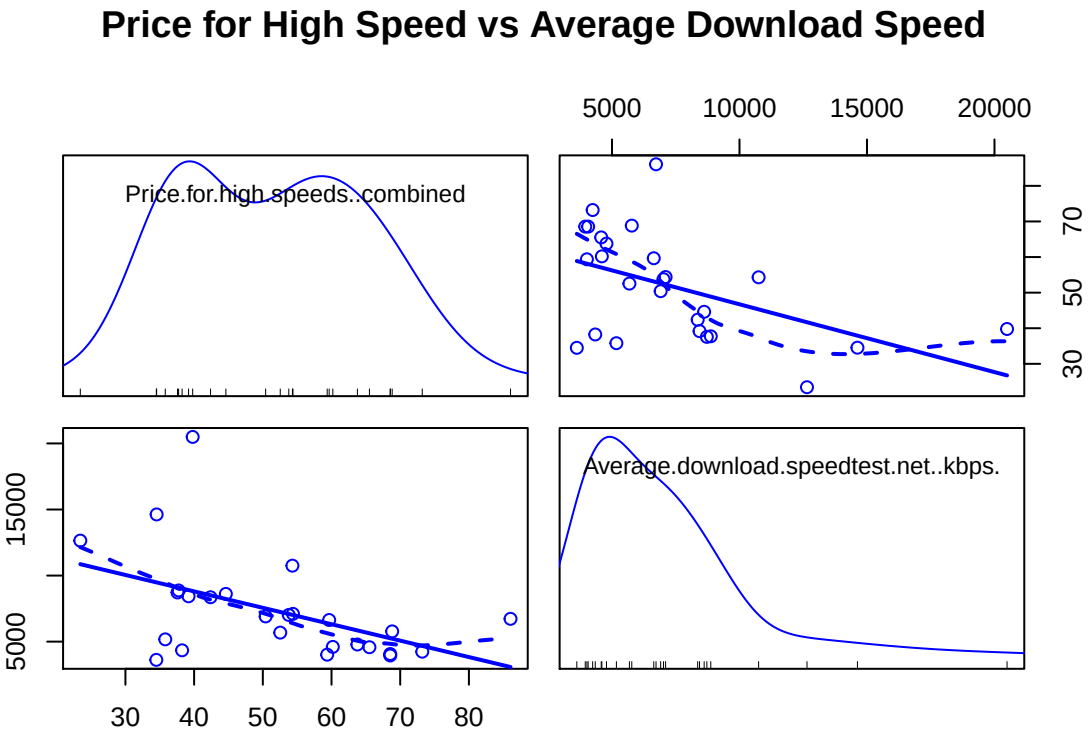
1. From the box plot for medium speed, we can see that the penetration is higher for lower prices.
2. The penetration at price range $81-90 is slightly higher than at $71-80 for medium speeds, indicating that more people are willing to pay the higher price. But we don't have any data to indicate if this can be attributed to better speed, a more reliable network or any other factor.
3. Among medium speeds, penetration is the highest in price ranges $21-50
4. From the box plot for high speed, we can see that there is more penetration in the price ranges $31-70. There is an overall negative relationship between price and penetration where we can see as price increases penetration decreases.

We also attempt to compare penetration and price across the same sets of price intervals for both medium and high speed networks. The penetration is higher in medium speed networks at $21-30. However for all other price bins penetration is higher on higher speed networks than medium speed networks. This indicates that for the same price more people are subscribing to higher speeds than medium speeds.

Speed vs. Price

The third bivariate relationship to explore is one between speed and price. Initially we decide to create a simple 2 x 2 scatterplot matrix to see if it will provide useful insights into how these variables interact.

```
library(car)
price_for_high_speed_trim = subset(Broadband_Prep, `Price for high speeds, combined` > 10 &
                                     `Price for high speeds, combined` < 100 )
scatterplotMatrix( ~ `Price for high speeds, combined` +
                     `Average download speedtest.net (kbps)`,
                      data = price_for_high_speed_trim,
                    main="Price for High Speed vs Average Download Speed", smooth=list(spread=FALSE))
```



We find that indeed there are some useful insights that we can take away from observing the speed vs. price relationship. It is important to note that we have suppressed the min and max outliers in the price for high speed variable we are using. As a reminder we know that there are two missing data points and two outliers removed from the high price variable as discussed earlier in the report.

However all this considered there is definite negative correlation between price and speed. Thinking about the data intuitively one hypothesis for this finding may be that countries with higher average download speeds are also those countries that have better broadband infrastructure and can offer lower prices. In order to confirm this hypothesis we would need to investigate the data points further and supplement it with information about the state of broadband technology in those countries as well as potentially statistical modeling.

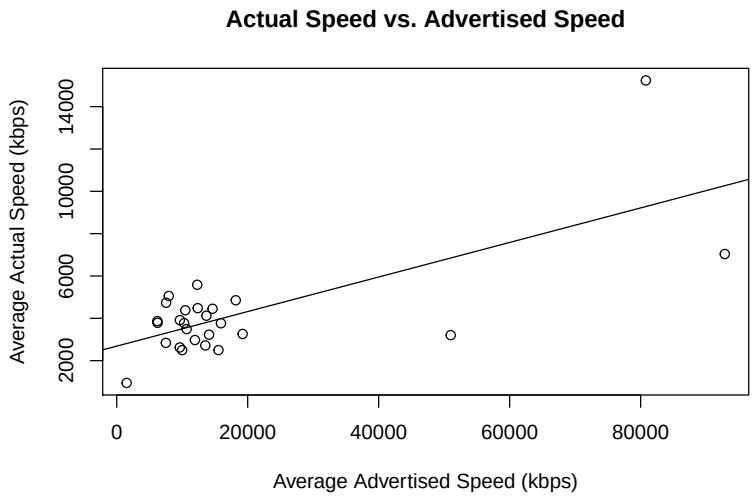
Analysis of Secondary Effects

Advertised vs. Actual Speeds

One secondary factor we wanted to explore in the broadband dataset was the relationship between advertised and actual speeds. We feel while this wasn't one of the main relationship between speed, penetration and

price it is relevant as it relates to the price of data rates that people are expecting to receive when the pay for higher data speeds.

```
plot(Broadband_Prep$`Average advertised speed OECD (kbps)`,
      Broadband_Prep$`Average actual speed, Akamai (kbps)`,
     main = "Actual Speed vs. Advertised Speed",
     xlab = "Average Advertised Speed (kbps)",
     ylab= "Average Actual Speed (kbps)")
abline(lm(Broadband_Prep$`Average actual speed, Akamai (kbps)` ~
            Broadband_Prep$`Average advertised speed OECD (kbps)`))
```




```
cor(Broadband_Prep$`Average actual speed, Akamai (kbps)`,
            Broadband_Prep$`Average advertised speed OECD (kbps)`,
            use = "complete.obs")
```

```
## [1] 0.704136
```

The actual vs. advertised speeds differ by an order of magnitude which seems to indicate that the internet speed people are paying for is a farce. However there is a strong correlation between advertised speed and actual speed, which does indicate that paying for higher speed connection seems to result in a higher speed connection, just not as high as what you might expect.

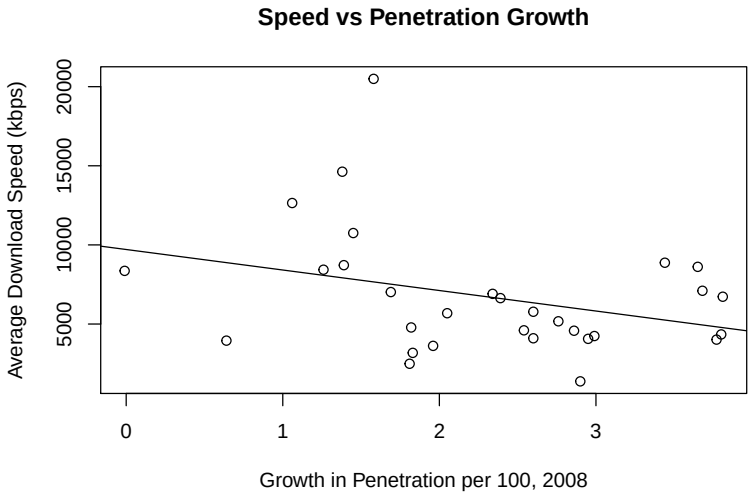
Other Penetration Variables

There were several interesting penetration variables that we investigated against speed and price. These were: - growth in penetration per 100 population from 2007 to 2008 - percentage of urban population

Speed vs Penetration Growth

First we plot growth in penetration per 100 vs average download speed. We see that they seem to be negatively related.

```
plot(growth, Broadband_Prep$`Average download speedtest.net (kbps)`,
     main="Speed vs Penetration Growth",
     ylab="Average Download Speed (kbps)", xlab="Growth in Penetration per 100, 2008")
abline(lm(Broadband_Prep$`Average download speedtest.net (kbps)` ~ growth))
```



We get a numerical measure of this correlation. There is a negative correlation of -0.32.

```
cor(growth, Broadband_Prep$`Average download speedtest.net (kbps)`)
```

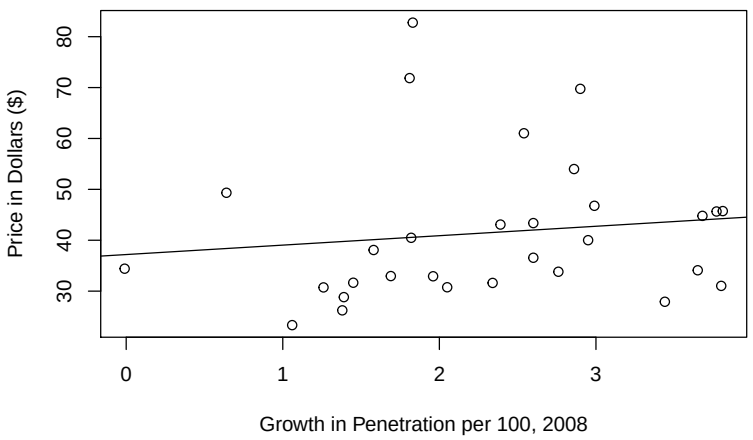
```
## [1] -0.3274673
```

The faster the growth in penetration, the lower the average download speed found in that country. A possible reason for this may be that countries with high speeds already have high penetration rates, so further growth in penetration is harder. As a result, growth in penetration in these countries may be slower.

Prices vs Penetration Growth

Next, we examine the relationship between prices and penetration growth. We plot these variables to find how they are related.

```
plot(growth, Broadband_Prep$`Price for med speeds, combined`,
     main="Price for Medium Speeds vs Penetration Growth", ylab="Price in Dollars ($)",
     xlab="Growth in Penetration per 100, 2008")
abline(lm(Broadband_Prep$`Price for med speeds, combined` ~ growth))
```



We get a numerical measure of this correlation. There is very little correlation, at 0.13. We can say that these variables are unrelated.

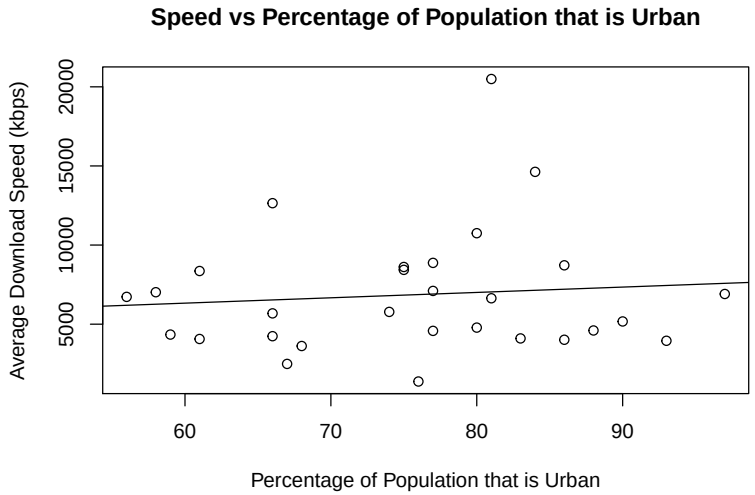
```
cor(growth, Broadband_Prep$`Price for med speeds, combined`)
```

```
## [1] 0.1290671
```

Speed vs Percentage of Population that is Urban

We plot speed vs percentage of population that is urban. There is an outlier of 160% of people in a country being in urban areas that is removed for this plot.

```
plot(Broadband_Prep$`Percent of population in urban areas`[Broadband_Prep$`Percent of population in urba
abline(lm(Broadband_Prep$`Average download speedtest.net (kbps)`[Broadband_Prep$`Percent of population i
```



We get a numerical measure of this correlation. There is a basically no correlation, with a value of 0.09.

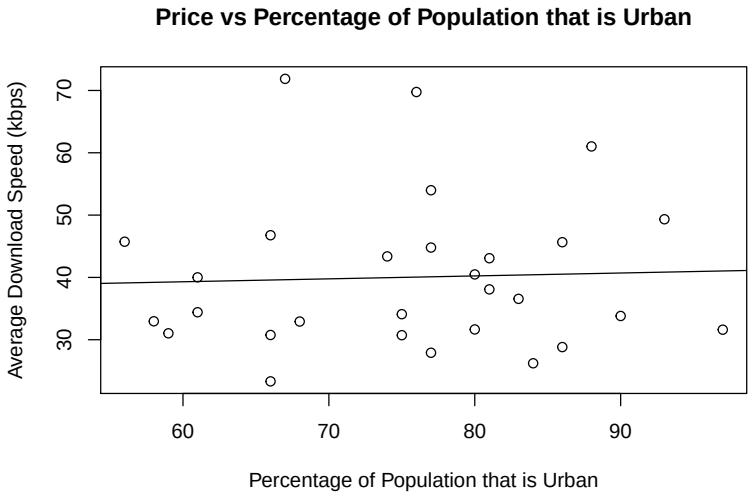
```
cor(Broadband_Prep$`Percent of population in urban areas`[Broadband_Prep$`Percent of population in urban
```

```
## [1] 0.09421763
```

Price vs Percentage of Population that is Urban

We plot price vs percentage of population that is urban. We continue to remove the outlier data point where 160% of a country's population is in urban areas.

```
plot(Broadband_Prep$`Percent of population in urban areas`[Broadband_Prep$`Percent of population in urba
abline(lm(Broadband_Prep$`Price for med speeds, combined`[Broadband_Prep$`Percent of population in urban
```



We get a numerical measure of this correlation. There is also basically no correlation, with a value of 0.04.

```
cor(Broadband_Prep$`Percent of population in urban areas`[Broadband_Prep$`Percent of population in urba
```

```
## [1] 0.04231162
```

We can conclude from these analyses that urbanization is unrelated to both measures of speed and price.

Conclusion

The exploratory analysis of the given broadband data has helped us see how speed, penetration and price relate with one another. Based on our analysis we can simply summarize that as speed increases, penetration increases and as price increases, penetration drops. We also observed that as speed increases prices drop, which was originally a counter intuitive finding.

We also share that countries like the United States, Slovak Republic and Mexico do not gain any speed, price or penetration advantage by not adopting open access regulations and this may even served as a disadvantage based on our analysis. They have download speeds lesser than the average download speeds, have less than the median penetration rate and have prices more than the average prices when compared to countries that have adopted some form of open access regulations. We feel confident in our finding that there is evidence that may require further statistical modeling but that these countries should adopt open access regulation.

Finally, we explored whether other factors related to penetration (growth in penetration from 2007 to 2008 as well as percent of population in urban areas) were related to speed or prices. While there was some negative correlation between the amount of growth in penetration from 2007 to 2008 and speed, but the percent of population in urban areas was not strongly correlated with Internet speeds or prices.