An exploratory analysis of Broadband Data

Dili Wang, Kaitlin Swinnerton, Sartaj Singh Baveja

9/23/2018

Introduction:

A 2010 study by the Berkman Center for Internet and Society reported that the majority of the OECD (Organisation for Economic Cooperation and Development) Countries had adopted some form of open access policies, which required broadband owners to share part of their infrastructure with other companies at regulated rates to prevent monopoly pricing or to promote an increase in penetration. However, 3 countries - the United States, Mexico, and the Slovak Republic - were discovered as exceptions to having adopted open access. Network owners in the the US, in particular, argued that unregulated pricing, or facility based competition, incentivizes investment in better quality equipment and promotes penetration to hard to reach rural areas. Ultimately, they believe that there is a trade-off between price, speed and penetration.

In this exploratory analysis, we will be studying broadband data for 30 countries, consisting of 3 major variable categories - Price, Penetration, and Speed. We are motivated by the following questions:
1. Main: In investigating the three-way relationship between price, speed and penetration, Does a trade-off exist between these concepts?
2. Secondary: Is there evidence for the beneficial effects of open access policies?

Outline of Analysis

1. In univariate analysis, we will closely study the variables that pertain to each of the major concepts, Price, Penetration and Speed. In this part of the analysis, we will look at key features in summary statistics, identify potenital outliers or erroneous data, and examine how multiple variables within one category, which measure the same concept, compare with each other. This will help drive our consideration for which relationships between concepts are considered essential.
2. Using the key variables from univariate analysis, we will examine all possible bivariate relationships between the concepts. In the relationships analysis, we will examine Penetration vs Price, Penetration vs Speed, and Speed vs Price. Lastly, we will discuss any relationships that we can see between all 3.
3. In the analysis of Secondary Effects, we will examine if countries that do not have open access policies exhibit any key features for variables within each of the 3 categories, as well as any of the bivariate relationships. We may also identify other variables, both internal and external to the data set, that could potentially confound our understanding of the data or the relationships that we have identified.
4. In the conclusion, we will summarize our analysis results and synthesize our answers to both the main and secondary research questions.

Loading the Data:

The data is first loaded into 3 separate dataframes, named after their respective files. The individual data frames will be cleaned, to be used for univariate analysis. They will also be joined, post cleaning, into one dataframe later on, to be used for relationship analysis

```
setwd("~/Box\ Sync/w203/lab_1/broadband/Final\ Report")
getwd()
```

```
## [1] "/Users/diliwang/Box Sync/w203/lab_1/broadband/Final Report"
```

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

We will also be using the following libraries throughout our analysis:

```
library(ggrepel)
library(dplyr)
library(car)
library(reshape2)
library(ggplot2)
library(cowplot)
```

For some initial observation, we first look at the number of rows and columns of each of the 3 dataframes:

```
nrow(Speed)
```

```
## [1] 32
```

```
length(names(Speed))
```

```
## [1] 17
```

```
nrow(Penetration)
```

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

```
length(names(Penetration))
```

```
## [1] 13
```

```
nrow(Price)
```

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

```
length(names(Price))
```

```
## [1] 6
```

We also took a look at the column headers to see if there are columns that all of the dataframes share, which can be used as a key for merging the dataframes

```
#output intentionally supressed here, as we will discuss the columns in great detail later.
names(Speed)
names(Penetration)
names(Price)
```

We see here that all 3 dataframes share the columns “Country” and “Country.code”. Besides these columns which appear as text, the remaining columns of each dataframe appear to be numeric metrics. In summary, we first see that:

1. The Speed dataframe contains 2 text variables and 15 metric variables. There are 31 observations total.
2. The Penetration dataframe contains the same 2 text variables and 11 metric variables. However, the 13th column strangely has the header “X”, which indicates to us that the raw data has no header for the 13th column. There are a total of 32 observations
3. The Speed dataframe contains 2 text variables and 4 metric variables. There are a total of 31 observations.

Besides the appearance of one column with header “X”, the most alarming feature we noticce in the raw data is that the individual dataframes do not have the same number of observations upon the initial dataload. This must be resolved prior to merging of the dataframes into one. In observing the csv files prior to load, the each of the 3 files show 30 rows, so the only dataframe that currently has the correct number of observations is the Price dataframe. The Speed dataframe, which has 32 observations, appears to have 2 extra rows, while the Penetration with 31 rows appears to have 1 extra row.

To close examine the extra rows, we print them in our rmd. We can confirm from this output that the 31st row of the Penetration dataframe contains only NA or blank values. Similarly, the 31st and 32nd rows of the Speed dataframe contains only NA or blank

```
Penetration[31,]
Speed[c(31, 32),]
```

With this evidence as justifiation that these rows do not contain data, we decide to remove these empty rows from their respective dataframes prior to analysis or merging

```
Penetration = Penetration[-31, ]
Speed = Speed[-c(31,32), ]
```

As noted in the above, we also noticed a column in the Penetration dataframe with the header “X”, which we believed to represent an empty header. To check that this column is blank:

```
is.na(Penetration$X)
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE
```

Again, we see that this columns is blank, so we decide to remove the column “X” from the Penetrtion dataframe.

```
Penetration = Penetration[,-13]
```

Also from the outputs of our initial observations, we see that the “Country” and “Country.Code” headers are shared by every file. Therefore, we can use “Country” column as the key for table joining. In order to be a valid key for table joining in the dplyr library, we must ensure that the “Country” columns are of class character:

```
Speed$Country <- as.character(Speed$Country)
Penetration$Country <- as.character(Penetration$Country)
Price$Country <- as.character(Price$Country)
```

We also confirmed that the “Country” columns have the same values for each of the 3 data frames:

```
Speed$Country == Penetration$Country
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE
```

```
Penetration$Country == Price$Country
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE
```

Additional Data Cleaning:

Before we continue with dataframe merging and analysis, it would make sense to see if any of the numeric data requires additional cleaning. Let’s first start with a summary of the Price dataframe:

```
summary(Price)
```

Currently, the 4 currency columns of the Price dataframe are all datatype factor with arbitrarily assigned levels. This is likely due to the pre-existing of the formatting of the data, which includes “$” that prevents the data from being treated as numeric when read into a dataframe. In order to perform proper analysis on Price as well as its relationship with other variables, we need to turn the columns of the price dataframe into numeric vectors. We will perform the processing method:

1. removes starting “$” symbol
2. transforms all of the remaining characters after the “$” symbol to numbers using as.numeric()

```
convert_to_num_price = function(x) as.numeric(gsub("\\$", "", x))
Price$Price.for.low.speeds..combined <- convert_to_num_price(Price$Price.for.low.speeds..combined)
Price$Price.for.med.speeds..combined <- convert_to_num_price(Price$Price.for.med.speeds..combined)
Price$Price.for.high.speeds..combined <- convert_to_num_price(Price$Price.for.high.speeds..combined)
Price$Price.for.very.high.speeds..combined <- convert_to_num_price(Price$Price.for.very.high.speeds..combined)
```

Next, let’s take a look at the Penetration data frame to see if any additional transformations are needed:

```
summary(Penetration)
```

From our output in rmd, it appears that most of the columns of the Penetration dataframe is numeric. However, there are 2 colums which appear as factors rather than numeric due to the existence of pre-existing formatting with “%” symbol, that prevented the data from being treated as fully numeric when initially read into the dataframe.
Using another method (similar to Price), these “%” columns will need to be process into numeric columns before performing further analysis. To transform the columns, we will use a function that does the following:

1. remove the trailing “%” symbol
2. transforms all remaining characters to numbers using as.numeric()

```
convert_to_num_pen = function(x) as.numeric(gsub("%", "", x))
Penetration$Growth.in.3G.penetration <- convert_to_num_pen(Penetration$Growth.in.3G.penetration)
Penetration$Percent.of.population.in.urban.areas <- convert_to_num_pen(Penetration$Percent.of.population.in.urba
n.areas)
```

Lastly, we will examine the Speed dataframe for potential variables that require cleaning:

```
summary(Speed)
```

From our output in rmd, it appears that many of the columns in the Speed dataframe appear as factors, rather than numeric vectors, due to the presence of a “,” symbol in the pre-existing formatting. They will need to be transformed to numeric columns using a function that:

1. removes the “,” symbol
2. transforms all remaining characters to numbers using as.numeric()

```
convert_to_num_speed = function(x) as.numeric(gsub(",", "", x))
Speed$Maximum.advertised.speed.OECD..kbps. <- convert_to_num_speed(Speed$Maximum.advertised.speed.OECD..kbps.)
Speed$Average.advertised.speed.OECD..kbps. <- convert_to_num_speed(Speed$Average.advertised.speed.OECD..kbps.)
Speed$Average.actual.speed..Akamai..kbps. <- convert_to_num_speed(Speed$Average.actual.speed..Akamai..kbps.)
Speed$Average.download.speedtest.net..kbps. <- convert_to_num_speed(Speed$Average.download.speedtest.net..kbps.)
Speed$Standard.deviation.download..speedtest.net <- convert_to_num_speed(Speed$Standard.deviation.download..speed
test.net)
Speed$Average.upload.speedtest.net..kbps. <- convert_to_num_speed(Speed$Average.upload.speedtest.net..kbps.)
Speed$Standard.deviation.upload..speedtest.net <- convert_to_num_speed(Speed$Standard.deviation.upload..speedtes
t.net)
Speed$Standard.deviation.latency..speedtest.net <- convert_to_num_speed(Speed$Standard.deviation.latency..speedte
st.net)
Speed$Median.download..speedtest.net..kbps. <- convert_to_num_speed(Speed$Median.download..speedtest.net..kbps)
Speed$Median.upload..speedtest.net..kbps. <- convert_to_num_speed(Speed$Median.upload..speedtest.net..kbps.)
Speed$X90p..Download..speedtest.net..kbps. <- convert_to_num_speed(Speed$X90p..Download..speedtest.net..kbps.)
Speed$X90p..Upload..speedtest.net..kbps. <- convert_to_num_speed(Speed$X90p..Upload..speedtest.net..kbps)
```

Post cleanig, we are now ready to initialize merging. Since all dataframes have the exact same number of rows and exactly identical “Country” values, we can use full_join. First we merge Price and Penetration dataframes

```
Price_Penetration = full_join(Price, Penetration, by = "Country")
```

Then we merge Price_Penetration with Speed dataframe to form the Full_Data dataframe, consisting of all 3 individual dataframes

```
Full_Data = full_join(Price_Penetration, Speed, by  = "Country")
```

Lastly we check all of the column headers to make sure that the column headers in Full_Data include all of the column headers of the Speed, Penetration, and Price dataframes.

```
names(Full_Data)
```

Once all of the column headers are confirmed, in Full_Data, we notice that the “Country.Code” columns are repeated twice, as expected. We then remove the extra ones as they are not needed.

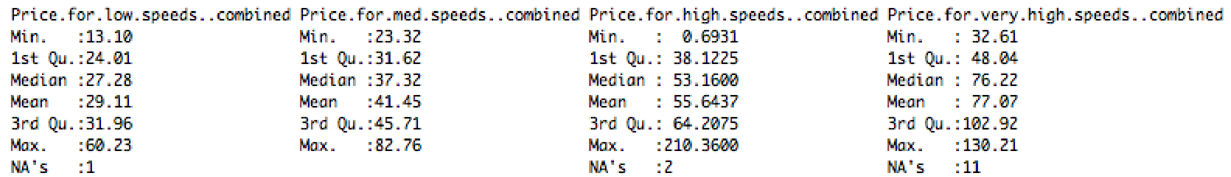
```
Full_Data = Full_Data[, -c(7, 18)]
names(Full_Data)[2] = 'Country.Code'
```

Univariate Analysis of Key Variables

Price Variable

First, I start with summary statistics on the numeric columns in the original Price dataframe. Since the Price dataframe was cleaned prior to merging, these summary statistics are also reflected in the Full_data dataframe.

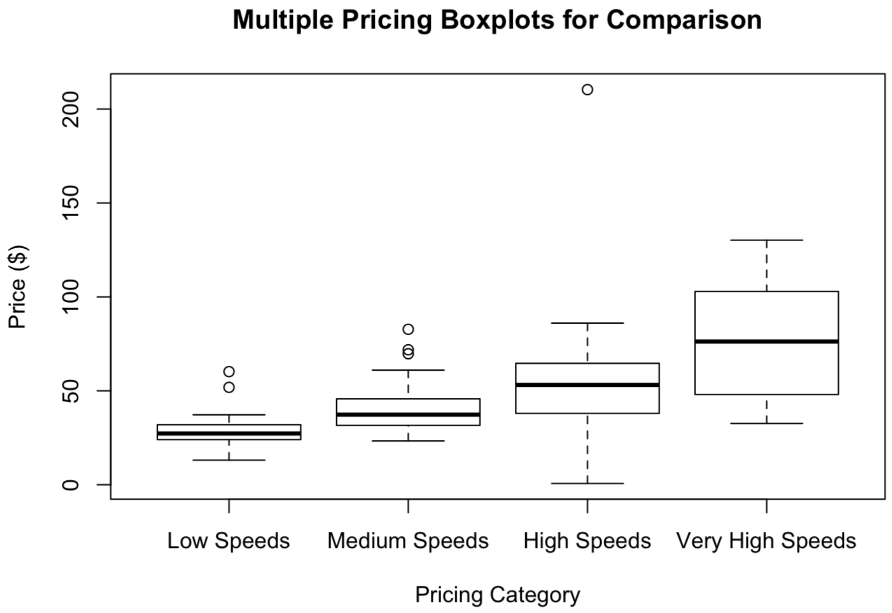
```
summary(Price[c(3,4,5,6)])
```



```
Price.for.low.speeds..combined Price.for.med.speeds..combined Price.for.high.speeds..combined Price.for.very.high.speeds..combined
Min.   :13.10                  Min.   :23.32                  Min.   :  0.6931                Min.   : 32.61                     
1st Qu.:24.01                  1st Qu.:31.62                  1st Qu.: 38.1225                1st Qu.: 48.04                     
Median :27.28                  Median :37.32                  Median : 53.1600                Median : 76.22                     
Mean   :29.11                  Mean   :41.45                  Mean   : 55.6437                Mean   : 77.07                     
3rd Qu.:31.96                  3rd Qu.:45.71                  3rd Qu.: 64.2075                3rd Qu.:102.92                     
Max.   :60.23                  Max.   :82.76                  Max.   :210.3600                Max.   :130.21                     
NA's   :1                                                     NA's   :2                       NA's   :11                         
```

To supplement the comparison of each of the 4 numeric columns, this aggregate boxplot figure is also included:

```
boxplot(Price$Price.for.low.speeds..combined, Price$Price.for.med.speeds..combined,
        Price$Price.for.high.speeds..combined, Price$Price.for.very.high.speeds..combined,
        main = "Multiple Pricing Boxplots for Comparison",
        xlab = "Pricing Category", ylab = "Price ($)",
        names = c("Low Speeds", "Medium Speeds", "High Speeds", "Very High Speeds"))
```



A few key observations:

1. There are 4 numeric Price variables, distinguished by speed levels - Low, Medium, High, Very High
2. Only the Medium speed column has complete data, while the other 3 columns are incomplete. Low and Medium speed columns contain a few NA values, while Very High Speed column has the most NA values for 11/30 observations. Our initial conclusion here is that the missing values are not necessarily an indicator that the data is poor quality. Rather, this could mean that many countries, approximately 1/3 in this data set, possibly do not have the infrastructure capabilities to acheived speeds that are categorized as “Very High”
3. We also see here, as expected, that the Median and Mean Prices increases as the Pricing Category increases in speed from Low to Very High.
4. In addition, we can notice here that both the Range and the IQR of each column overlaps with the next column that indicates one speed higher in the Pricing Category. For instance, range and IQR of Low Speeds prices overlaps with that of Medium Speeds prices. We then see the same trend for Medium and High, High and Very High.

Exceptions/Anomalies:

1. For the Low Speeds pricing category, we see extreme values at $51.86 (Australia) and $60.23 (Mexico), which are potential outliers. However, since Mexico is a key country that does not have open access policy, we do not feel like there is sufficient justification here for removal, and have decided to keep both values.
2. Extreme values for Medium Speeds pricing category include $69.76 (Mexico), $77.86 (Turkey), $82.76 (Poland), but they are not far enough away from IQR to warrant removal.
3. High Speeds pricing category has 2 very extreme values. The first is a maximum at $210.36 (Poland). It is unclear if this value is erroneous, and in fact could be potentially accurate given the fact that Poland already reports the highest price for the Medium speeds pricing category one tier below. The High Speeds Price for Poland also exceeds all prices in the very high category. We will make a note of this for bivariate analysis later.
4. High Speeds pricing category also has one of the lowest prices observed, the value $0.6931 (Luxembourg). This definitely appears to be an erroneous value for 2 reasons. The first is that this is the only value with 4 decimal spaces in the entire Price dataframe, which leads us to suspect that the input is user error and that the correct value should be $69.31. Secondly, all other countries show an increase in pricing as the Pricing category increases in speed. Luxembourg with this extremely low value is an anomaly, unless we change the price to $69.31. Therefore, we believe we are justified in modifying this value in both the Price and Full_Data dataframes:

```
Full_Data$Price.for.high.speeds..combined[Full_Data[[2]] == 'LU'] = 69.31
Price$Price.for.high.speeds..combined[Price[[2]] == 'LU'] = 69.31
```

5. Additionally, we are listing here all of the countries that are missing values for any of the columns, in case they affect bivariate analysis later. These include:

```
Price$Country[is.na(Price$Price.for.low.speeds..combined)]
```

```
## [1] "Belgium"
```

```
Price$Country[is.na(Price$Price.for.high.speeds..combined)]
```

```
## [1] "Mexico" "Turkey"
```

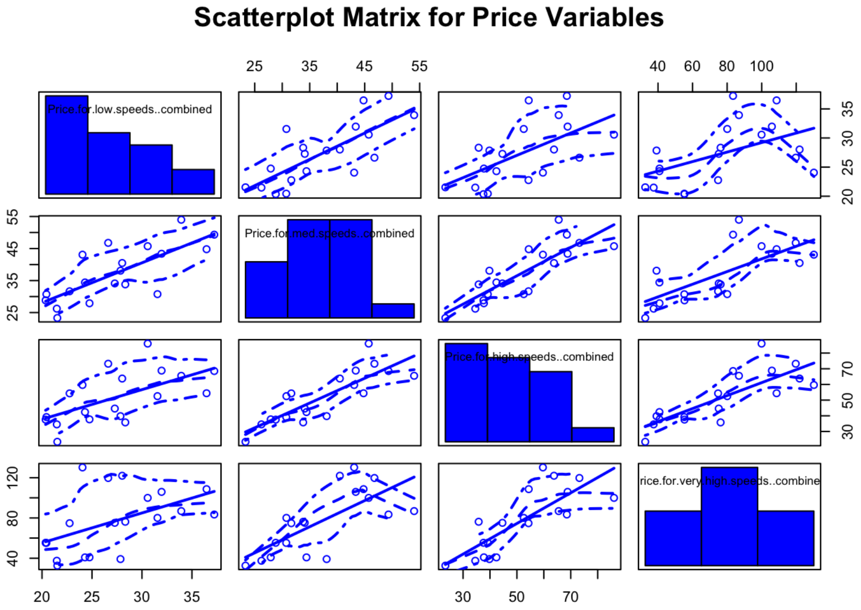
```
Price$Country[is.na(Price$Price.for.very.high.speeds..combined)]
```

```
##  [1] "Australia"   "Belgium"     "Greece"      "Ireland"     "Italy"
##  [6] "Luxembourg"  "Mexico"      "New Zealand" "Poland"      "Portugal"
## [11] "Turkey"
```

From this output, we notice all of the countries which have Very High speeds pricing has values for Low, Medium and High Speeds pricing, but the opposite is not true. This could be further evidence that the missing price values for Very High speeds are not actually due to poor data quality, but are the result infrastructure limitations within these countries. A couple countries have neither High Speeds or Very High Speeds pricing (Mexico and Turkey), while Belgium is the only country to have no Low Speeds pricing.

For a sanity check, we also plotted the Price columns against each other in a scatterplot matrix:

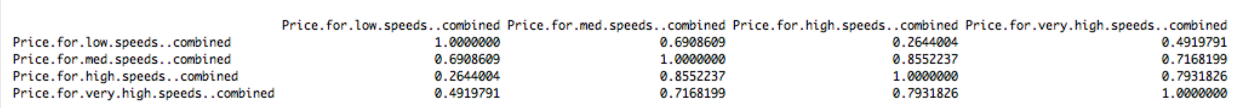
```
scatterplotMatrix( Price[seq(3,6)],
                   main = "Scatterplot Matrix for Price Variables", diagonal=list(method='histogram'))
```


Scatterplot Matrix for Price Variables

This is mainly to check

1. If there are any surprises in the data, such as negative correlation between any 2 pricing column. Luckily, we see that all correlations between Price columns are positive as we would expect
2. The correlations between consecutive columns appear to be stronger, than the correlation between non-consecutive columns (say Low Speeds vs Very High Speeds, for instance)

```
cor(Price[seq(3,6,1)], use = 'pairwise.complete.obs')[,c(1,2,3,4)]
```



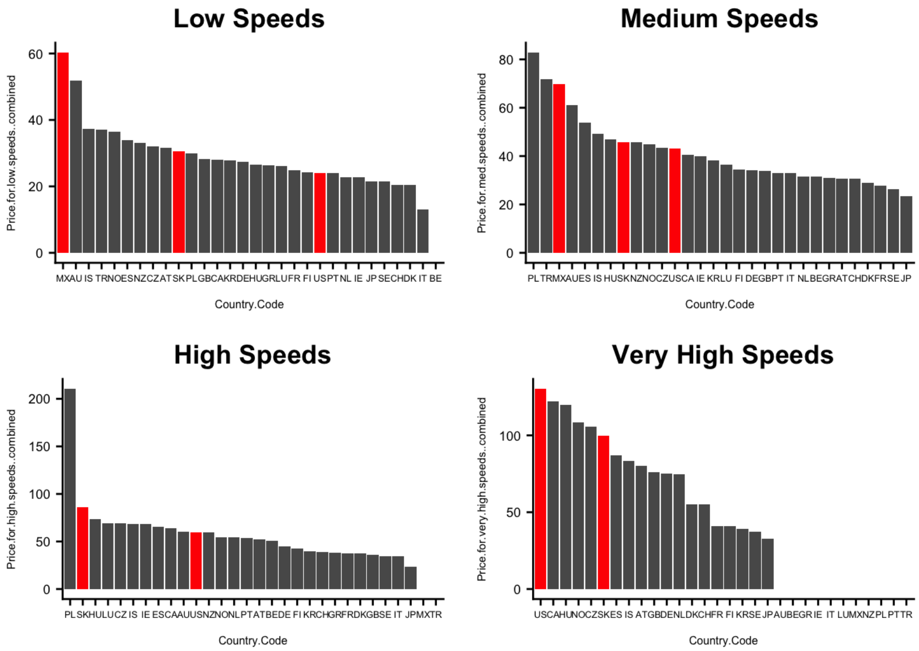
```
                                     Price.for.low.speeds..combined Price.for.med.speeds..combined Price.for.high.speeds..combined Price.for.very.high.speeds..combined
Price.for.low.speeds..combined                            1.0000000                      0.6908609                       0.2644004                            0.4919791
Price.for.med.speeds..combined                            0.6908609                      1.0000000                       0.8552237                            0.7168199
Price.for.high.speeds..combined                           0.2644004                      0.8552237                       1.0000000                            0.7931826
Price.for.very.high.speeds..combined                      0.4919791                      0.7168199                       0.7931826                            1.0000000
```

Some limitations with the data to consider: 1) We have no background within the data on how the Pricing Levels are determined, as in which speeds constitute the definitions for Low, Medium, High, and Very High. As a result, we will not be able to combine these columns into one column using a reasonable transformation. 2) The IQR and range of the Very High Speeds pricing category is very large, and the number of observations for this column is also quite low. This is something for us to consider further when studying the validity of bivariate relationships

In summary, it looks like the Medium Speeds category and the High Speeds Category are very highly correlated. Given that the Medium Speeds category is most complete and has the least extreme outlier of the 2 columns, the Medium Speeds Price variable is the best candidate for close examination of bivariate and/or 3 way relationships later on.

Countries who did not adopt Open Access Lastly, in preparation to analyze if there are any benefits to open access policies, we can plot each of the Prices within one categry as a function of the Country, while highlighting the countries that did not adopt open access policies:

```
# Low Speed Prices per Country, ordered by Low Speed Prices
Price$Country.Code <- factor(Price$Country.Code, levels = Price$Country.Code[order(-Price$Price.for.low.speeds..c
ombined)])
low = Price %>%
  mutate(highlight_flag = ifelse(Country.Code == 'US' | Country.Code == 'MX' | Country.Code == 'SK', T, F)) %>%
  ggplot(aes(x = Country.Code, y = Price.for.low.speeds..combined)) +
  ggtitle("Low Speeds")+
  geom_bar(stat = "identity", aes(fill = highlight_flag), position = "dodge", show.legend = FALSE) +
  theme(legend.justification=c(1,1),legend.position=c(1,1),legend.title=element_blank(), text=element_text(size=6
), axis.text.x=element_text(size=5), axis.text.y=element_text(size=7))+
  scale_fill_manual(values = c('#595959', 'red'))
# Medium Speed Prices per Country, ordered by Medium Speed Prices
Price$Country.Code <- factor(Price$Country.Code, levels = Price$Country.Code[order(-Price$Price.for.med.speeds..c
ombined)])
med = Price %>%
  mutate(highlight_flag = ifelse(Country.Code == 'US' | Country.Code == 'MX' | Country.Code == 'SK', T, F)) %>%
  ggplot(aes(x = Country.Code, y = Price.for.med.speeds..combined)) +
  ggtitle("Medium Speeds")+
  geom_bar(stat = "identity", aes(fill = highlight_flag), position = "dodge", show.legend = FALSE) +
  theme(legend.justification=c(1,1),legend.position=c(1,1),legend.title=element_blank(), text=element_text(size=6
), axis.text.x=element_text(size=5), axis.text.y=element_text(size=7))+
  scale_fill_manual(values = c('#595959', 'red'))
# High Speed Prices per Country, ordered by High Speed Prices
Price$Country.Code <- factor(Price$Country.Code, levels = Price$Country.Code[order(-Price$Price.for.high.speeds..
combined)])
high = Price %>%
  mutate(highlight_flag = ifelse(Country.Code == 'US' | Country.Code == 'MX' | Country.Code == 'SK', T, F)) %>%
  ggplot(aes(x = Country.Code, y = Price.for.high.speeds..combined)) +
  ggtitle("High Speeds")+
  geom_bar(stat = "identity", aes(fill = highlight_flag), position = "dodge", show.legend = FALSE) +
  theme(legend.justification=c(1,1),legend.position=c(1,1),legend.title=element_blank(), text=element_text(size=6
), axis.text.x =element_text(size=5), axis.text.y=element_text(size=7))+
  scale_fill_manual(values = c('#595959', 'red'))
# Very High Speed Prices per Country, ordered by Very High Speed Prices
Price$Country.Code <- factor(Price$Country.Code, levels = Price$Country.Code[order(-Price$Price.for.very.high.spe
eds..combined)])
vhigh = Price %>%
  mutate(highlight_flag = ifelse(Country.Code == 'US' | Country.Code == 'MX' | Country.Code == 'SK', T, F)) %>%
  ggplot(aes(x = Country.Code, y = Price.for.very.high.speeds..combined)) +
  ggtitle("Very High Speeds")+
  geom_bar(stat = "identity", aes(fill = highlight_flag), position = "dodge", show.legend = FALSE) +
  theme(legend.justification=c(1,1),legend.position=c(1,1),legend.title=element_blank(), text=element_text(size=6
), axis.text.x =element_text(size=5), axis.text.y=element_text(size=7))+
  scale_fill_manual(values = c('#595959', 'red'))
plot_grid(low, med, high, vhigh, scale = 1, label_size = 8)
```



From these plots, we can draw a few conclusions:

1. For Low Speeds Pricing Category, Mexico is the leading in the highest speed, closely followed by Australia. All other countries have lower speeds, with Slovak Republic close to the center and United States closer to the low end of pricing.

2. For Medium Speeds, Poland, Turkey and once again, Mexico, lead in the highest prices, while Slovak Republic and US are middle of the range in pricing.
3. For High Speeds, with the exception of the Poland's price outlier, the Slovak republic is now a leader in the highest price. The US is still in center of the pricing range, while Mexico has no data point for this speed category.
4. For Very High Speeds, the United States is a clear leader in price, with Slovak republic close to the center of the range, and Mexico, again, with no data point.
5. Japan and Sweden show consistently some of the lowest prices across all categories of speed.

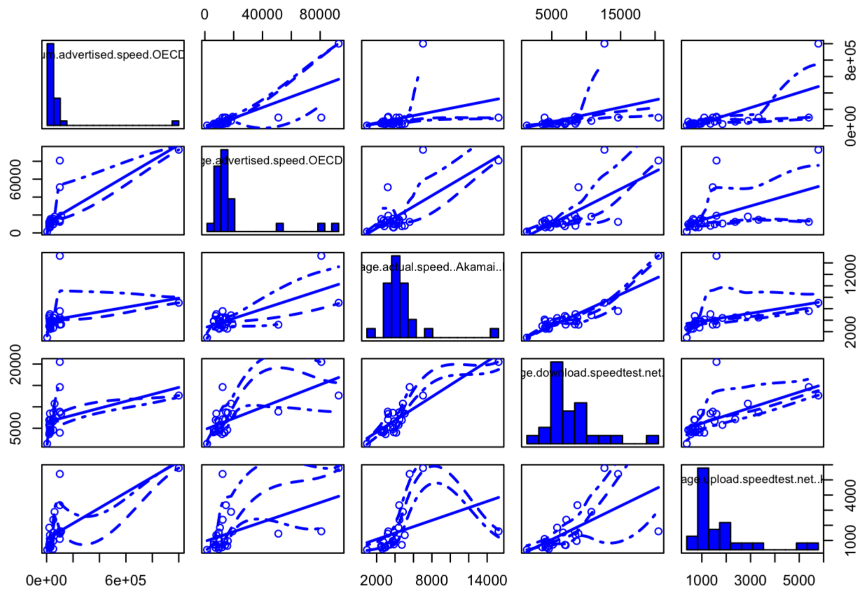
From the single variate analysis, we see that one of the 3 countries that has does not have open access policies are leaders in pricing for each of the speed-based pricing categories. Mexico has the hightest price for Low speeds, one of the highest prices of Medium speeds; the Slovak Republic has one of the highest prices for high speeds; the US has the highest price for Very High Speeds. Because these countries have higher pricing for at least one of these speed categories compared to other countries, this may indicate a possible disadvantage in having not adopted open access policies.

Speed Variable

To begin with, we start off the analysis with a scatterplot matrix. This is helpful for getting a high-level overview of the relationships between our variables and can draw attention to important features we want to investigate further.

We start by looking at the different average and maximum speeds observed by OECD, Akamai and Speedtest.

```
scatterplotMatrix(~ Maximum.advertised.speed.OECD..kbps. + Average.advertised.speed.OECD..kbps. +  Average.actua
l.speed..Akamai..kbps. + Average.download.speedtest.net..kbps. + Average.upload.speedtest.net..kbps., data=Full_D
ata, diagonal=list(method="histogram"))
```



We notice a couple of features in this matrix that can help guide our analysis.

1. There is a noticable positive relationship between average advertised speeds measured by the OECD to the average actual speeds measured by Akamai. The advertised speeds are consistently higher than actual speeds and this makes sense since latency is added in as a factor. Advertised speeds do in fact offer a reasonable prediction of the variation across countries in actual speeds.
2. Another thing to notice from the above is that, that the Average Actual Speed measured by Akamai is highly correlated with the Average Download Speed measured by Speedtest. This happens even though there are 4 NA's in the speeds measured by Akamei.

To look further into this,

```
summary(Full_Data$Average.actual.speed..Akamai..kbps.)
```

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

```
summary(Full_Data$Average.download.speedtest.net..kbps.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1377    4135    5730    6729    8415   20493
```

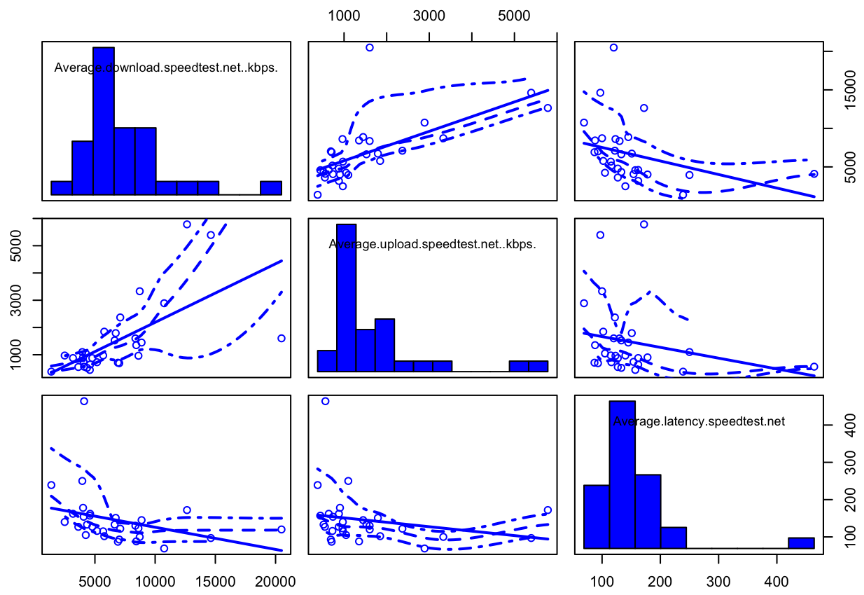
From the summary, it can be seen that both of the speed measures have a few outliers on the extreme end that could be driving this relationship. In order to see what the relationship looks like without those outliers, we'll plot the relationships excluding the outliers.

```
scatterplotMatrix(~ Average.actual.speed..Akamai..kbps. + Average.download.speedtest.net..kbps.,
                  data = Full_Data[Full_Data$Average.actual.speed..Akamai..kbps. < 10000 &
                                   Full_Data$Average.download.speedtest.net..kbps. < 15000, ],
                  diagonal=list(method="histogram"))
```

The positive relationships persist, so we can conclude that these relationships are not driven by the outliers, and will continue our analyses with the outliers included.

Next, let's take a look at the Average Latency measured by speedtest.net. Theoretically, latency measures the time it takes to establish a connection between two systems. As the latency increases, the download and upload speeds should decrease.

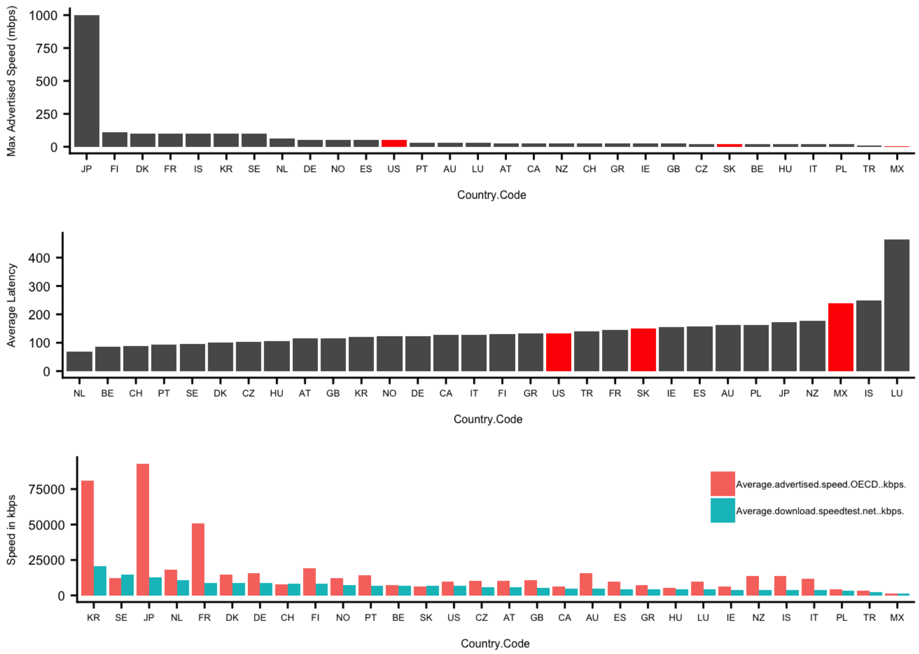
```
scatterplotMatrix(~ Average.download.speedtest.net..kbps. + Average.upload.speedtest.net..kbps. + Average.latenc
y.speedtest.net, data=Full_Data, diagonal=list(method="histogram"))
```



No surprises there. We see a negative relationship among latency vs average upload and download speeds.

For the following bar chart plots, we try to analyze the countries US, Mexico and Slovak Republic where there is no form of open access, compared against the other countries in the dataset.

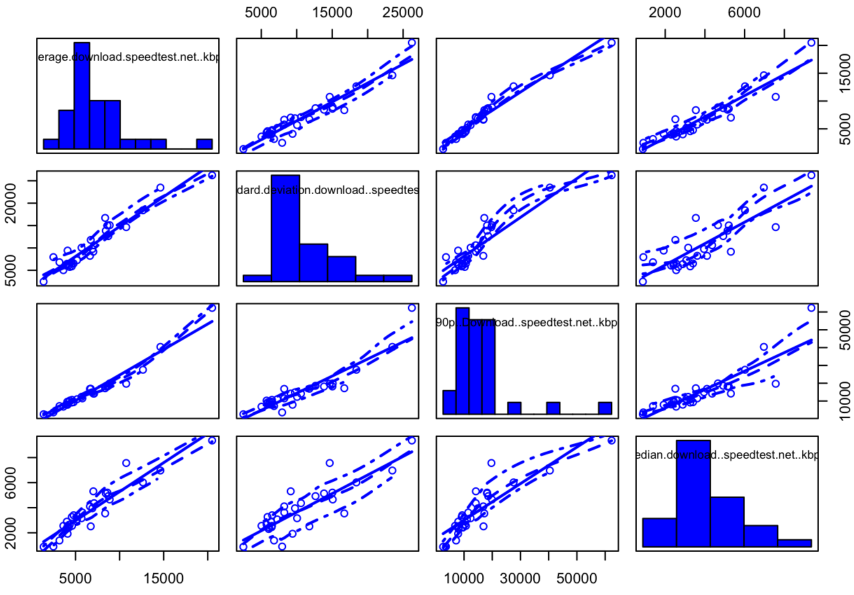
```
# Maximum Advertised vs Country Code
Full_Data$Country.Code <- factor(Full_Data$Country.Code, levels = Full_Data$Country.Code[order(-Full_Data$Maximu
m.advertised.speed.OECD..kbps.)])

p1 = Full_Data %>%
  mutate(highlight_flag = ifelse(Country.Code == 'US' | Country.Code == 'MX' | Country.Code == 'SK', T, F)) %>%
  ggplot(aes(x = Country.Code, y = Maximum.advertised.speed.OECD..kbps./1000)) + geom_bar(stat = "identity", aes
(fill = highlight_flag), show.legend=F) + theme(text=element_text(size=6), axis.text.x=element_text(size=5), axi
s.text.y=element_text(size=7)) + scale_fill_manual(values = c('#595959', 'red')) + ylab("Max Advertised Speed (mb
ps)")

# Average Latency vs Country Code
Full_Data$Country.Code <- factor(Full_Data$Country.Code, levels = Full_Data$Country.Code[order(Full_Data$Average.
latency.speedtest.net)])

p2 = Full_Data %>%
  mutate(highlight_flag = ifelse(Country.Code == 'US' | Country.Code == 'MX' | Country.Code == 'SK', T, F)) %>%
  ggplot(aes(x = Country.Code, y = Average.latency.speedtest.net)) + geom_bar(stat = "identity", aes(fill = highl
ight_flag), show.legend=F) + theme(text=element_text(size=6), axis.text.x=element_text(size=5), axis.text.y=eleme
nt_text(size=7)) + scale_fill_manual(values = c('#595959', 'red')) + ylab("Average Latency")

Full_Data$Country.Code <- factor(Full_Data$Country.Code, levels = Full_Data$Country.Code[order(-Full_Data$Averag
e.download.speedtest.net..kbps.)])
Full_Data_tf <- melt(Full_Data[, c("Country.Code", "Average.advertised.speed.OECD..kbps.", "Average.download.spee
dtest.net..kbps.")], id.vars = 1)

p3 = Full_Data_tf %>%
  ggplot(aes(x = Country.Code, y = value)) + geom_bar(stat = "identity", aes(fill = variable), position = "dodge"
) + theme(legend.justification=c(1,1),legend.position=c(1,1),legend.title=element_blank(), text=element_text(size
=6), axis.text.x=element_text(size=5), axis.text.y=element_text(size=7)) + ylab("Speed in kbps")

plot_grid(p1, p2, p3, nrow = 3, ncol = 1)
```



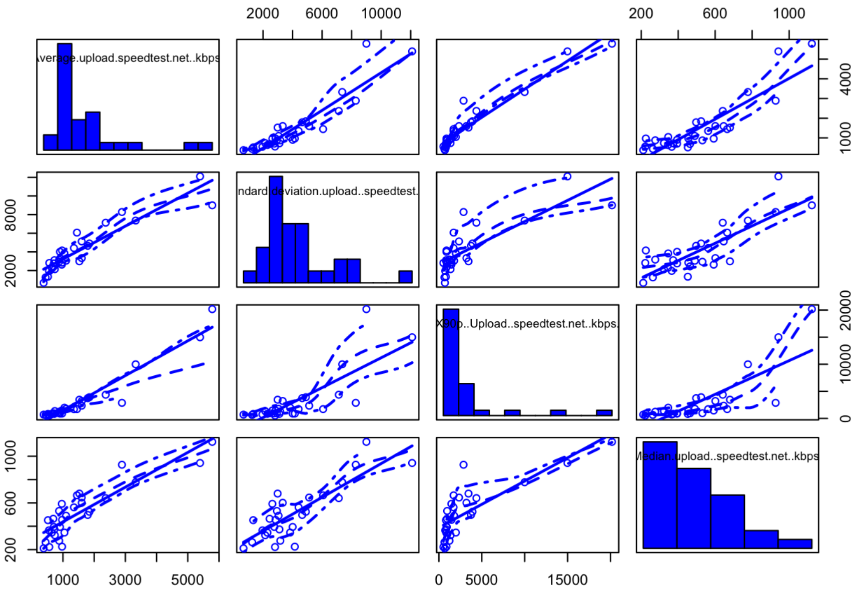
From the above plots, we notice that:
1. Looking at the Maximum Advertised Speed, Japan has the fastest speeds. In comparison, US is lagging behind comparable countries such as Japan, Korea, Finland and France for instance.
2. Similarly, for latency, Mexico has a pretty high latency as measured by speedtest. Thus, this proves why it has very low speeds.
3.There is a massive difference in average advertised speed and average download speed in Japan, Korea, France, Finland etc. In the case of US, it does better on average download speed than average advertised speed.

Finally, we aim to see how do the plots comparing Average, top 10% and Standard Deviation values fare against median values observed for download speeds, upload speeds and latency.

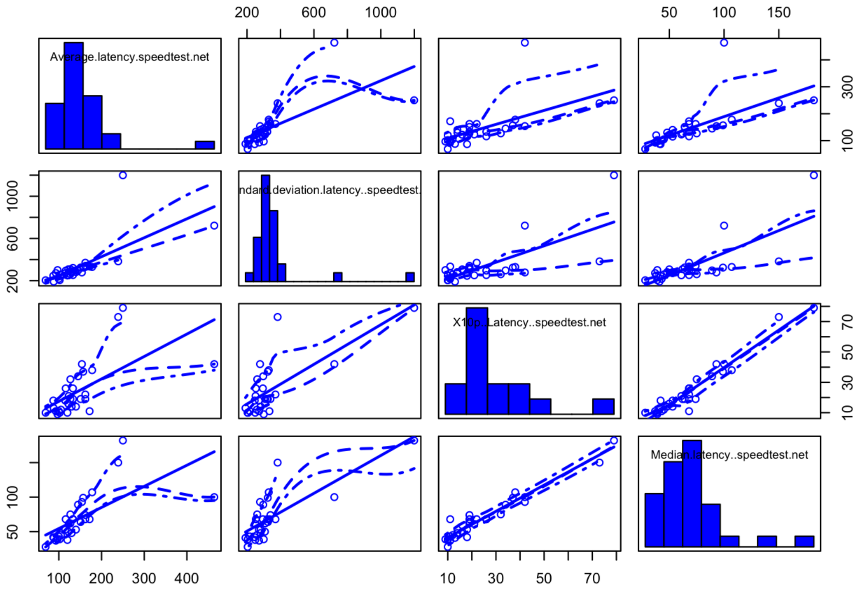
```
scatterplotMatrix(~ Average.download.speedtest.net..kbps. + Standard.deviation.download..speedtest.net + X90p..Do
wnload..speedtest.net..kbps. + Median.download..speedtest.net..kbps., data=Full_Data, diagonal=list(method="histo
gram"))
```



```
scatterplotMatrix(~ Average.upload.speedtest.net..kbps. + Standard.deviation.upload..speedtest.net + X90p..Uploa
d..speedtest.net..kbps. + Median.upload..speedtest.net..kbps., data=Full_Data, diagonal=list(method="histogram"))
```



```
scatterplotMatrix(~ Average.latency.speedtest.net + Standard.deviation.latency..speedtest.net + X10p..Latency..sp
eedtest.net + Median.latency..speedtest.net, data=Full_Data, diagonal=list(method="histogram"))
```



The above series of correlation graphs offer us some degree of confidence. As these graphs show, for the download speed, upload speed and latency, average measurements are well correlated with median measurements which in turn are well correlated with top 10% of measurements. In all the cases, the results are well spread out for download and upload speeds. As for latency, the data is much more noisier.

Penetration Variable

Let's look at the penetration measures.

```
summary(select(Full_Data, Country, Country.Code, Penetration.per.100.OECD..2008,
                       Penetration.per.100.OECD..2007, Household.penetration..OECD,
                       X2G.and.3G.penetration.per.100..OECD, Penetration.per.100.GC, X3G.penetration.per.100,
                       Growth.in.3G.penetration, Wi.Fi.hotspots..JiWire, Wi.Fi.hotspots.per.100.000..JiWire,
                       Percent.of.population.in.urban.areas))
```

```
   Country           Country.Code Penetration.per.100.OECD..2008 Penetration.per.100.OECD..2007 Household.penetration..OECD X2G.and.3G.penetration.per.100..OECD
 Length:30          AT     : 1    Min.   : 7.20                  Min.   : 4.30                  Min.   : 1.73               Min.   : 62.11                      
 Class :character   AU     : 1    1st Qu.:17.68                  1st Qu.:15.24                  1st Qu.:30.50               1st Qu.: 97.77                      
 Mode  :character   BE     : 1    Median :25.60                  Median :23.14                  Median :50.19               Median :110.00                      
                    CA     : 1    Mean   :23.96                  Mean   :21.66                  Mean   :46.41               Mean   :108.26                      
                    CH     : 1    3rd Qu.:30.51                  3rd Qu.:29.68                  3rd Qu.:62.97               3rd Qu.:117.79                      
                    CZ     : 1    Max.   :37.18                  Max.   :35.79                  Max.   :94.13               Max.   :151.39                      
                    (Other):24                                                                                                                                  
 Penetration.per.100.GC X3G.penetration.per.100 Growth.in.3G.penetration Wi.Fi.hotspots..JiWire Wi.Fi.hotspots.per.100.000..JiWire Percent.of.population.in.urban.areas
 Min.   : 6.30          Min.   : 0.00           Min.   :  0.00           Min.   :    6          Min.   : 0.600                     Min.   :56.00                       
 1st Qu.:16.35          1st Qu.:12.42           1st Qu.: 43.98           1st Qu.:  525          1st Qu.: 4.397                     1st Qu.:66.00                       
 Median :25.80          Median :26.28           Median : 58.20           Median : 2336          Median :13.255                     Median :77.00                       
 Mean   :23.19          Mean   :27.18           Mean   : 97.70           Mean   : 6899          Mean   :18.984                     Mean   :75.45                       
 3rd Qu.:30.38          3rd Qu.:37.74           3rd Qu.:113.75           3rd Qu.: 5286          3rd Qu.:22.227                     3rd Qu.:83.00                       
 Max.   :36.90          Max.   :71.80           Max.   :510.90           Max.   :67718          Max.   :74.270                     Max.   :97.00                       
                                                                                                                                   NA's   :1                           
```

So just by looking at the summary statistics, a few things pop out. First, the maximum value for "Percent.of.population.in.urban.areas" is 162. Since it's not possible to have over 100% of your population living in urban areas, this value must be a mistake. We will exclude this value from future analyses.

```
Full_Data$Percent.of.population.in.urban.areas[Full_Data$Percent.of.population.in.urban.areas > 100] <- NA
```

Additionally, we notice that both Growth.in.3G.penetration and Wi.Fi.hotspots..JiWire have very large ranges of values. For analyzing wifi hotspots, we can simply use the Wi.Fi.hotspots.per.100.000..JiWire measure instead. This measure is more meaningful since the values are normalized based on the population, meaning that the size of the country isn't confounding how we measure wifi penetration.

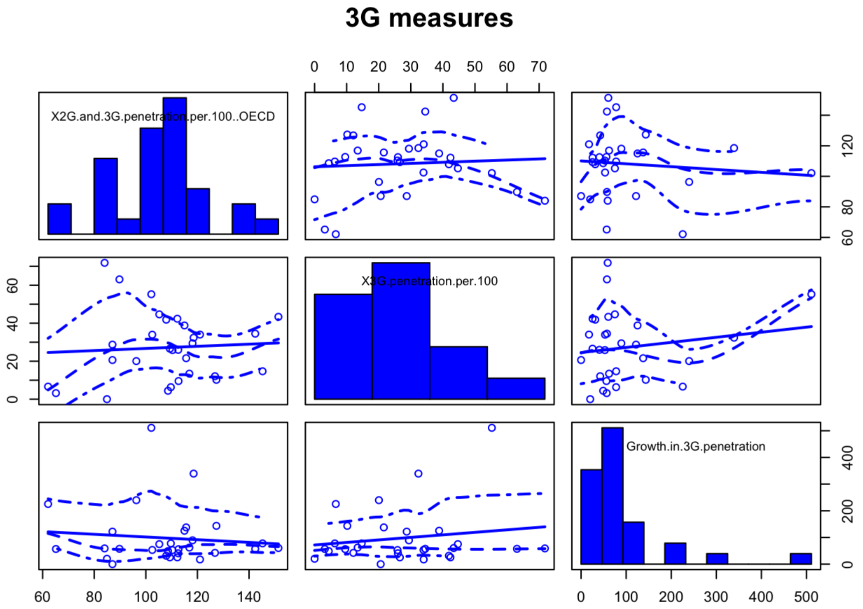
Additionally, it looks like we have some overlapping measures because our penetration measures come from two different data sets: The OECD Broadband Portal and the TeleGeography Global Comms Database. Next, we will check to see if these overlapping measures are in fact representing the same data.

First, we have three different measures of Penetration per 100: 1. Penetration per 100 OECD 2008 2. Penetration per 100 OECD 2007 3. Penetration per 100 CG (This data is from 2009)

These measures should theoretically be highly correlated. After plotting all three, we were able to confirm that they are in fact highly correlated. Knowing that they are all highly correlated, we will choose to only look at the penetration per 100 in 2008 for the remainder of the analyses. The other data we have ranges from 2007-2009, so we'll choose the 2008 data as it represents the median time point.

Next, we will look at our next set of overlapping data: X2G.and.3G.penetration.per.100..OECD and X3G.penetration.per.100 (from the TeleGeography Global Comms Database - henceforth referred to as GC). We will also look at how both relate to the Growth in 3G penetration from the GC database.

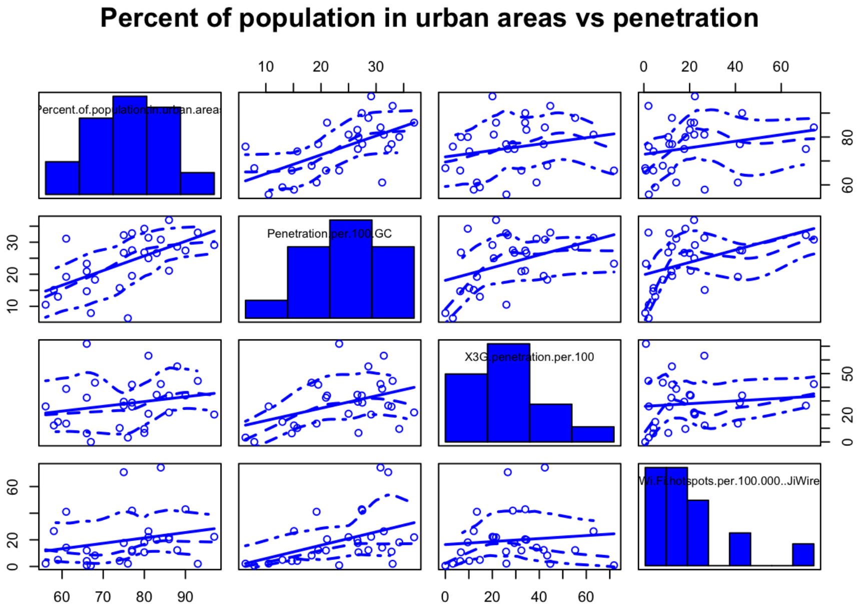
```
scatterplotMatrix(~ X2G.and.3G.penetration.per.100..OECD + X3G.penetration.per.100 +
                    Growth.in.3G.penetration,
                    data = Full_Data, diagonal=list(method="histogram"), main = "3G measures")
```



Surprisingly, we do not see any relationship between the X2G and 3G penetration measures from the OECD database and the 3G penetration measures from the GC database. It is possible that this is due to these measures being collected in different years. The GC databse is from 2009, but it is unlcear what year the OECD data is from. Additionally, growth in 3G penetration doesn't seem to have a strong relatioship with either 3G penetration measure.

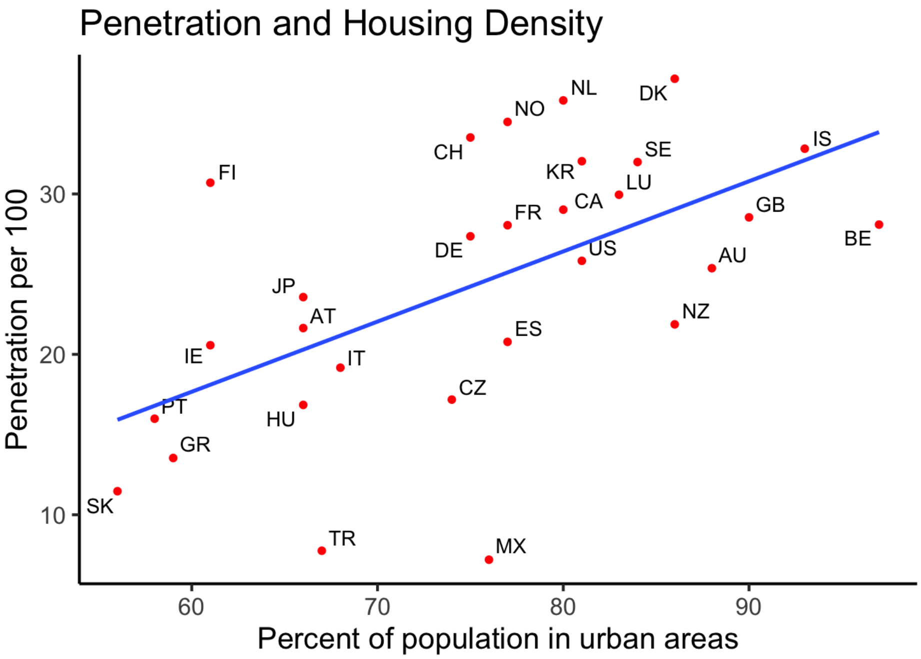
So now let's look at the three different measures of penetration and see how that relates to percent of the population in urban areas. Since the percent of population in urban areas comes from the CG database, we will use the penetration per 100 and 3G meaures from that database as well.

```
scatterplotMatrix(~ Percent.of.population.in.urban.areas + Penetration.per.100.GC  +
                    X3G.penetration.per.100 + Wi.Fi.hotspots.per.100.000..JiWire,
                    data = Full_Data, diagonal=list(method="histogram"), main="Percent of population in urban are
as vs penetration")
```



Percent of population in urban areas demonstrates positive relationships of varying degrees with each measure of penetration. It appears to have the strongest correlation with the penetration per 100 measure, such that countries that have a higher percentage of their population living in urban areas have a higher rate of penetration. This makes sense since the more densely populated a country is, the easier it is to achieve high rates of penetration. This is because providing broadband access to rural areas requires additional infrastructure.

```
ggplot(Full_Data, aes(Percent.of.population.in.urban.areas, Penetration.per.100.OECD..2008, label = Full_Data[[2
]])) +
  geom_text_repel() +
  geom_smooth(method = "lm", se = FALSE) +
  geom_point(color = 'red') +
  theme_classic(base_size = 16) +
  xlab("Percent of population in urban areas") +
  ylab("Penetration per 100")+
  ggtitle("Penetration and Housing Density")
```

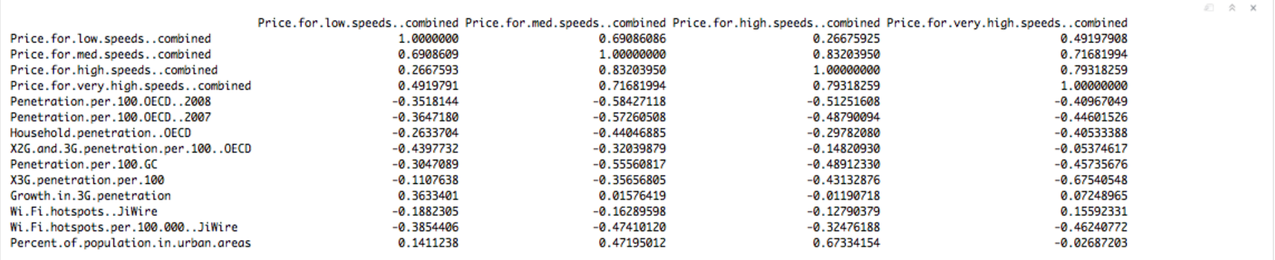


Analysis of Key Relationships

Price vs Penetration

To detect the potential bivariate relationships that warrant further study, we first display a subset of the Correlation Matrix that shows correlation coefficients for linear models between every Price variable and every Penetration variable.

```
#only enough of the correlation matrix is displayed here to see all bi-variate linear correlations
cor(Full_Data[seq(3,16,1)], use = 'pairwise.complete.obs')[,c(1,2,3,4)]
```



```
                                     Price.for.low.speeds..combined Price.for.med.speeds..combined Price.for.high.speeds..combined Price.for.very.high.speeds..combined
Price.for.low.speeds..combined                            1.0000000                     0.69086086                      0.26675925                           0.49197908
Price.for.med.speeds..combined                            0.6908609                     1.00000000                      0.83203950                           0.71681994
Price.for.high.speeds..combined                           0.2667593                     0.83203950                      1.00000000                           0.79318259
Price.for.very.high.speeds..combined                      0.4919791                     0.71681994                      0.79318259                           1.00000000
Penetration.per.100.OECD..2008                           -0.3518144                    -0.58427118                     -0.51251608                          -0.40967049
Penetration.per.100.OECD..2007                           -0.3647180                    -0.57260508                     -0.48790094                          -0.44601526
Household.penetration..OECD                              -0.2633704                    -0.44046885                     -0.29782080                          -0.40533388
X2G.and.3G.penetration.per.100..OECD                     -0.4397732                    -0.32039879                     -0.14820930                          -0.05374617
Penetration.per.100.GC                                   -0.3047089                    -0.55560817                     -0.48912330                          -0.45735676
X3G.penetration.per.100                                  -0.1107638                    -0.35656805                     -0.43132876                          -0.67540548
Growth.in.3G.penetration                                  0.3633401                     0.01576419                     -0.01190718                           0.07248965
Wi.Fi.hotspots..JiWire                                   -0.1882305                    -0.16289598                     -0.12790379                           0.15592331
Wi.Fi.hotspots.per.100.000..JiWire                       -0.3854406                    -0.47410120                     -0.32476188                          -0.46240772
Percent.of.population.in.urban.areas                      0.1411238                     0.47195012                      0.67334154                          -0.02687203
```

From the correlation matrix alone, with no outlier removal, we see that all of the relationships between Penetration and Price show negative correlation, for every Pricing level based on Speed, with the exception of the follwing:

1. “Growth in 3G Penetration” vs all Price variables
2. “Percent of Population in urban areas” vs all Pricing variables
3. “WiFi hotspots (JiWire)” vs “Price for Very High Speeds Combined”

For the first 2, “Growth in 3G Penetration” and “Pecent of Population in urban areas”, these are not actually metrics of Penetration, but might serve to inform Penetration itself. “WiFi hotspots (JiWire)”, as stated in the univariate analysis of penetration, is a raw count that is not actually normalized for each country, and would therefore not be a good candidate for bivariate analysis, since it could vary drastically based on country population and size. The correlation coefficients are also quite low against several of the Price variables, so these 3 will be excluded in the initial bivariate analysis

Of the remaining Penetration variables, only the following show consistently strong negative correlation against all 4 speed variables:

1. “Penetration per 100 (OECD 2008)”
2. “Penetration per 100 (OECD 2007)”
3. “Penetration per 100 (GC)”
4. “Household Penetration (OECD)”
5. “3G Penetration per 100”
6. “WiFi hotspots per 100,000 (JiWire)”

To further narrow down the relationships we wish to study between penetration and speed, we can actually combine some of the penetration variables, and selecting one of them as a proxy for relationship assessment From the Penetration univariate analysis, we know that there are strong correlations in the following cases:

1. “Penetration per 100 (OECD 2008)” vs “Penetration per 100 (OECD 2007)” vs “Penetration per 100 (GC)”. We see also that these 3 variables measure the same normalized metric, just that the first 2 are taken from separate years and the 3rd is taken from a different source. The values are also relatively similar in range, median, and mean and none of them appear to significantly scaled. Therefore, we can use the average of the 3 as a proxy for “Penetration per 100”, and investigate its relationship with Price.

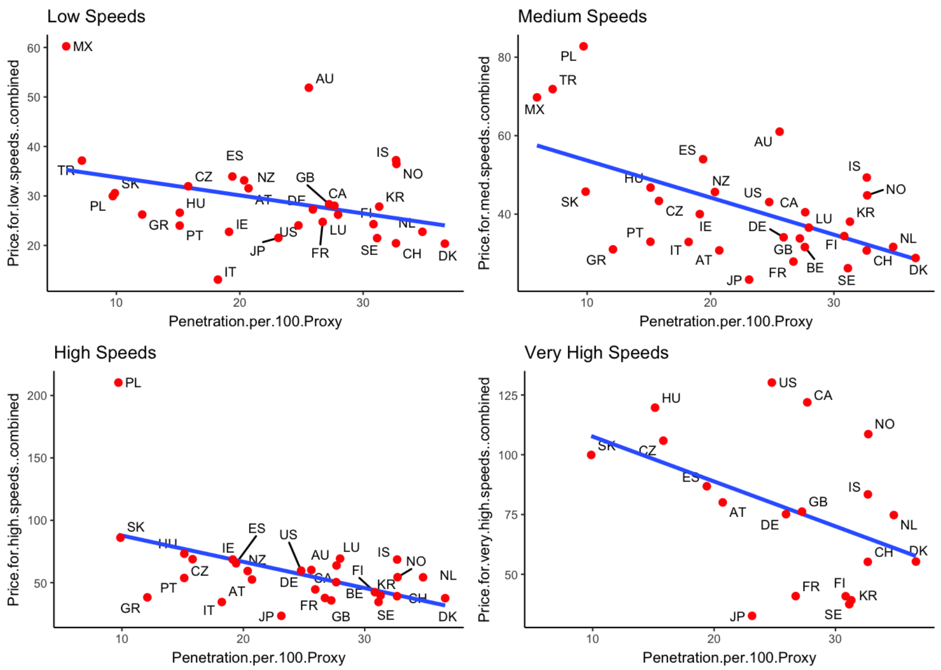
```
Full_Data["Penetration.per.100.Proxy"] = (Full_Data$Penetration.per.100.GC+Full_Data$Penetration.per.100.OECD..20
07+Full_Data$Penetration.per.100.OECD..2008)/3
```

2. “Household Penetration (OECD)” vs “3G Penetration per 100” is also strongly correlated in the univariate Penetration variables analysis. Again, we will choose a proxy to represent both variables in examining the relationship with Speed variables. The values in this two columns are not as closely related, so we cannot do an average for the proxy in this case. However, if we have to choose one of the 2, it should be “3G Penetration per 100” which has strong correlation for 3 of the 4 speed variables.

3. Lastly, we will also plot the normalized variable “WiFi hotspots per 100,000 (JiWire)” against the 4 price variables, given its relatively strong correlation coefficients.

1. “Penetration per 100 Proxy” vs Price

```
p1 = ggplot(Full_Data, aes( Penetration.per.100.Proxy, Price.for.low.speeds..combined, label = Full_Data[[2]])) +
 ggtitle("Low Speeds")+
  geom_text_repel(size=2.5) + geom_smooth(method = "lm", se = FALSE) + geom_point(color = 'red') + theme_classic
(base_size = 8)
p2 = ggplot(Full_Data, aes( Penetration.per.100.Proxy, Price.for.med.speeds..combined, label = Full_Data[[2]])) +
 ggtitle("Medium Speeds")+
  geom_text_repel(size=2.5) + geom_smooth(method = "lm", se = FALSE) + geom_point(color = 'red') + theme_classic
(base_size = 8)
p3 = ggplot(Full_Data, aes(Penetration.per.100.Proxy, Price.for.high.speeds..combined, label = Full_Data[[2]])) +
 ggtitle("High Speeds")+
  geom_text_repel(size=2.5) + geom_smooth(method = "lm", se = FALSE) + geom_point(color = 'red') + theme_classic
(base_size = 8)
p4 = ggplot(Full_Data, aes(Penetration.per.100.Proxy,Price.for.very.high.speeds..combined, label = Full_Data[[2
]])) + ggtitle("Very High Speeds")+
  geom_text_repel(size=2.5) + geom_smooth(method = "lm", se = FALSE) + geom_point(color = 'red') + theme_classic
(base_size = 8)
plot_grid(p1, p2, p3, p4, nrow = 2, ncol = 2)
```



First, from these plots, we see a clear negative correlation for each of the Price variables categorized by Speed tier vs the “Penetration per 100 Proxy” (the average of 2 Penetration per 100 columns). Even if we ignore some of the price outliers, say most notably Poland’s High Speed price of $210.36, we still see negative trend of “Penetration per 100” vs Price.

This outcome also seems reasonable with what we would expect logically from the consumer’s perspective. High penetration, or high percentage of population that are customer, is associated with lower Price, which is more favorable for the customer, while vice versa, Low penetration is associated with higher Price. The limitations of our data, however, does not allow us to discern causation between the 2 variables. For one, Price is also determined by the cost of the broadband infrasture as well as the possible competition between broadband companies in some countries. On the otherhand, Penetration also depend on factors like the percentage of population in rural areas as well as the geography and size of the country. Another limitation here is that we have no data on the percentage of penetration that subscribe to each of the speed tiers for pricing, which further confounds any attemtps at discerning causation.

The strength of the negative linear regression correlation between “Penetration per 100 Proxy” and Price does not appear to be affected by Pricing Tier in any way. From a graphical perspective, the correlation appears to be much stronger for Medium Speed prices and High Speed Prices as compared to the other 2 speed-based pricing tiers. It would also be expected that a correlation with Very High Speed prices is the lowest, given the wide range and sparce ness of the Very High Speeds pricing data. We can confirm our observations of correlation strength with a quick correlation coefficient output of the new column “Penetration per 100 Proxy” vs Price variables:

```
#only first column of the correlation matrix is needed here
cor(Full_Data[c(32,3,4,5,6)], use = 'pairwise.complete.obs')[, 1]
```

```
##            Penetration.per.100.Proxy       Price.for.low.speeds..combined 
##                            1.0000000                           -0.3417098 
##       Price.for.med.speeds..combined      Price.for.high.speeds..combined 
##                           -0.5724518                           -0.4841461 
## Price.for.very.high.speeds..combined 
##                           -0.4390432
```

An interesting phenomenon to investigate is the comparisons of slope of the line, used to fit linear regression, for each of 4 the “Penetration per 100 Proxy” vs Price variable relationships. We extract the slope coefficients from each of the 4 linear models

```
coef(lm(Full_Data$Price.for.low.speeds..combined~Full_Data$Penetration.per.100.Proxy))[2]
```

```
## Full_Data$Penetration.per.100.Proxy 
##                          -0.3656086
```

```
coef(lm(Full_Data$Price.for.med.speeds..combined~Full_Data$Penetration.per.100.Proxy))[2]
```

```
## Full_Data$Penetration.per.100.Proxy 
##                          -0.9441449
```

```
coef(lm(Full_Data$Price.for.high.speeds..combined~Full_Data$Penetration.per.100.Proxy))[2]
```

```
## Full_Data$Penetration.per.100.Proxy
##                           -2.103082
```

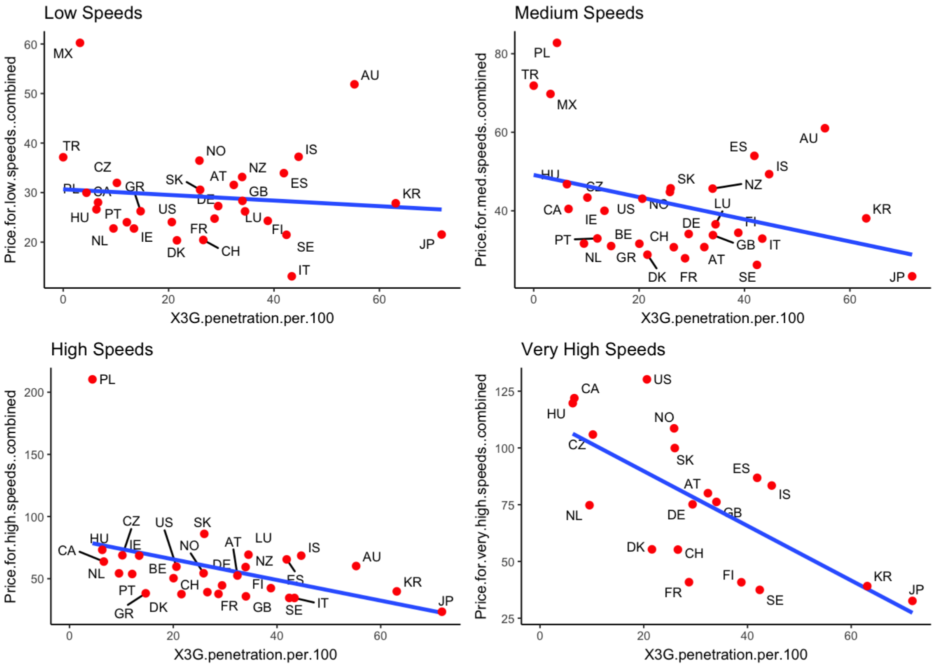
```
coef(lm(Full_Data$Price.for.very.high.speeds..combined~Full_Data$Penetration.per.100.Proxy))[2]
```

```
## Full_Data$Penetration.per.100.Proxy
##                           -1.878092
```

From this output, the low slope for the Low Price Tier relationship has already been indicated by the weakest correlation coefficient with “Penetration per 100 Proxy” from above. For the stronger bivariate correlations, we see a slope of ~-1 for the Penetration/100 vs Medium Speed Prices relationship, but a slope of ~2 for the Penetration/100 vs both High and Very High Speeds Prices relationships. Again, we are limited here by the lack of knowledge of weight of each of the Pricing Tiers (the percentage of customers that subscribe to each Speed based Pricing category). However, the 2x slope for High and Very High speeds compared to Medium Speeds could be an indication that pricing increase for Higher Speeds corresponds to a larger Penetration dropoff. This would seem reasonable, as customers who are price sensitive would be willing to settle for Medium speeds in the event that High and Very High Speed Tier’s price differences are too great.

2. “3G Penetration per 100” vs Price

```
p1 = ggplot(Full_Data, aes(X3G.penetration.per.100, Price.for.low.speeds..combined, label = Full_Data[[2]])) + gg
title("Low Speeds")+
  geom_text_repel(size=2.5) + geom_smooth(method = "lm", se = FALSE) + geom_point(color = 'red') + theme_classic
(base_size = 8)
p2 = ggplot(Full_Data, aes(X3G.penetration.per.100, Price.for.med.speeds..combined, label = Full_Data[[2]])) + gg
title("Medium Speeds")+
  geom_text_repel(size=2.5) + geom_smooth(method = "lm", se = FALSE) + geom_point(color = 'red') + theme_classic
(base_size = 8)
p3 = ggplot(Full_Data, aes(X3G.penetration.per.100, Price.for.high.speeds..combined, label = Full_Data[[2]])) + g
gtitle("High Speeds")+
  geom_text_repel(size=2.5) + geom_smooth(method = "lm", se = FALSE) + geom_point(color = 'red') + theme_classic
(base_size = 8)
p4 = ggplot(Full_Data, aes(X3G.penetration.per.100, Price.for.very.high.speeds..combined, label = Full_Data[[2
]])) + ggtitle("Very High Speeds")+
  geom_text_repel(size=2.5) + geom_smooth(method = "lm", se = FALSE) + geom_point(color = 'red') + theme_classic
(base_size = 8)
plot_grid(p1, p2, p3, p4, nrow = 2, ncol = 2)
```



From these graphical outputs, we see that the correlation strength between Price and “3G Penetration per 100” increases as the pricing tier speed becomes greater. From our intial correlation matrix output at the start of this section, the correlation coefficients are {-0.1107638 , -0.35656805 , -0.43132876 , -0.67540548} for {Low, Medium, High, Very High} Speed Prices vs “3G Penetration per 100”. An additional comparison on the slope of each line used for the 4 linear models show that slope decreases as the pricing tier speed increases.

```
#Extracting only the slope  from the linear models that fit each graph:
coef(lm(Full_Data$Price.for.low.speeds..combined~Full_Data$X3G.penetration.per.100))[2]
```

```
## Full_Data$X3G.penetration.per.100
##                         -0.0567869
```

```
coef(lm(Full_Data$Price.for.med.speeds..combined~Full_Data$X3G.penetration.per.100))[2]
```

```
## Full_Data$X3G.penetration.per.100
##                         -0.2825053
```

```
coef(lm(Full_Data$Price.for.high.speeds..combined~Full_Data$X3G.penetration.per.100))[2]
```

```
## Full_Data$X3G.penetration.per.100
##                         -0.8302984
```

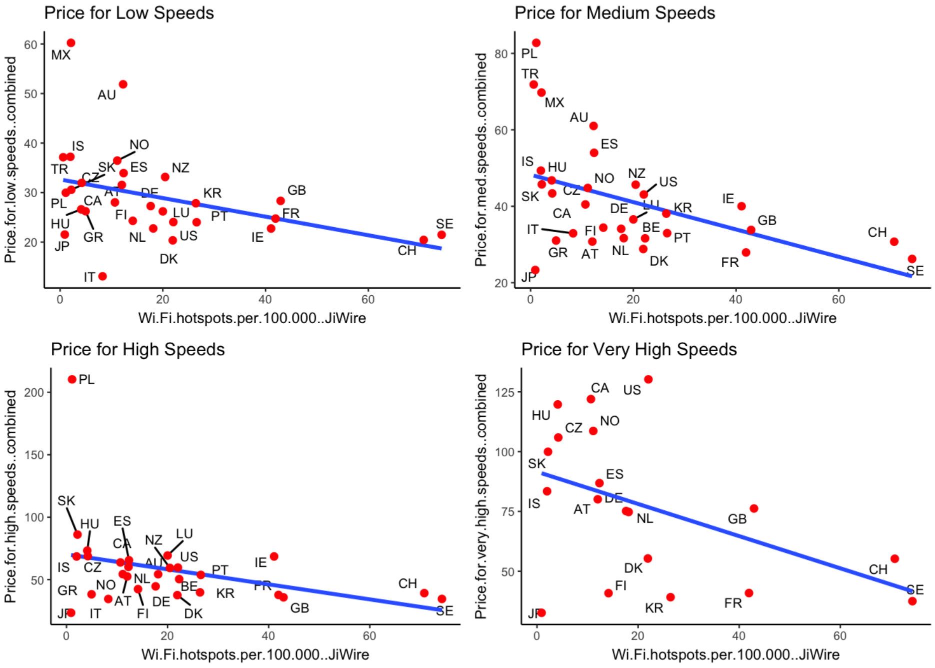
```
coef(lm(Full_Data$Price.for.very.high.speeds..combined~Full_Data$X3G.penetration.per.100))[2]
```

```
## Full_Data$X3G.penetration.per.100
##                          -1.204595
```

From the slope comparisons alone, it would seem that as Price increases for each speed tier, the drop off of penetration increases. This could indicate that price sensitive customers become less willing to pay for higher speed and higher priced 3G internet, and may be more willing to settle for the slower and less expensive 2G version.

3. “WiFi Hotspots per 100,000” vs Price

```
p1 = ggplot(Full_Data, aes(Wi.Fi.hotspots.per.100.000..JiWire, Price.for.low.speeds..combined, label = Full_Data
[[2]])) + ggtitle("Price for Low Speeds")+
  geom_text_repel(size=2.5) + geom_smooth(method = "lm", se = FALSE) + geom_point(color = 'red') + theme_classic
(base_size = 8)
p2 = ggplot(Full_Data, aes(Wi.Fi.hotspots.per.100.000..JiWire, Price.for.med.speeds..combined, label = Full_Data
[[2]])) + ggtitle("Price for Medium Speeds")+
  geom_text_repel(size=2.5) + geom_smooth(method = "lm", se = FALSE) + geom_point(color = 'red') + theme_classic
(base_size = 8)
p3 = ggplot(Full_Data, aes(Wi.Fi.hotspots.per.100.000..JiWire, Price.for.high.speeds..combined, label = Full_Data
[[2]])) + ggtitle("Price for High Speeds")+
  geom_text_repel(size=2.5) + geom_smooth(method = "lm", se = FALSE) + geom_point(color = 'red') + theme_classic
(base_size = 8)
p4 = ggplot(Full_Data, aes(Wi.Fi.hotspots.per.100.000..JiWire, Price.for.very.high.speeds..combined, label = Full
_Data[[2]])) + ggtitle("Price for Very High Speeds")+
  geom_text_repel(size=2.5) + geom_smooth(method = "lm", se = FALSE) + geom_point(color = 'red') + theme_classic
(base_size = 8)
plot_grid(p1, p2, p3, p4, nrow = 2, ncol = 2)
```



According to the earlier correlation matrix, the strongest correlation in the 4 relationships of vs Price variables vs “WiFi Hotspots per 100K” is actually exists for Medium Speeds prices and Very High Speeds Prices. This is not indicated in the graphs themselves, as Price for High Speeds vs “WiFi Hotspots per 100K” seems to have the strongest correlation graphically. However, upon closer examination, the linear models are all somewhat skewed by the “WiFi Hotspots per 100K” outliers for Switzerland and Sweden.

A comparison of linear model line slopes for each of the graphs shows some decrease in slopes as the Pricing Tier Speed increases:

```
coef(lm(Full_Data$Price.for.low.speeds..combined~Full_Data$Wi.Fi.hotspots.per.100.000..JiWire))[2]
```

```
## Full_Data$Wi.Fi.hotspots.per.100.000..JiWire
##                                   -0.1877494
```

```
coef(lm(Full_Data$Price.for.med.speeds..combined~Full_Data$Wi.Fi.hotspots.per.100.000..JiWire))[2]
```

```
## Full_Data$Wi.Fi.hotspots.per.100.000..JiWire
##                                   -0.3576902
```

```
coef(lm(Full_Data$Price.for.high.speeds..combined~Full_Data$Wi.Fi.hotspots.per.100.000..JiWire))[2]
```

```
## Full_Data$Wi.Fi.hotspots.per.100.000..JiWire
##                                   -0.6014605
```

```
coef(lm(Full_Data$Price.for.very.high.speeds..combined~Full_Data$Wi.Fi.hotspots.per.100.000..JiWire))[2]
```

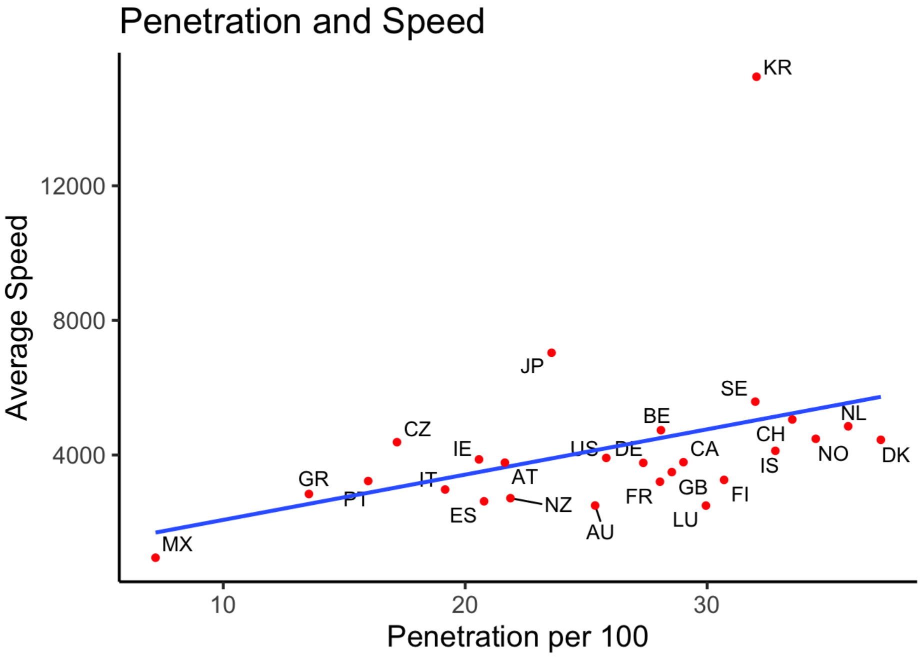
```
## Full_Data$Wi.Fi.hotspots.per.100.000..JiWire
##                                   -0.6727032
```

We can consider, to some extent, that “Wi.Fi.hotspots.per.100.000”, or the density of Hotspot locations amongst number of people, to be an measure of the strength of a country’s WiFi infrastrucure. This increase in Price drop-off corresponding to both an increase in pricing tier speed and increase in Hotspot density could indicate a couple of potential relationships. The first could be that the higher the hotspot density, the stronger the WiFi infrastructure, which could lead to lower cost and thus lower pricing. The second relationship could be that the higher the price, the lower the penetration, and thus the lower the investment potential to be able to increase hotspot density and WiFi infrastructure. These relationships are currently speculative and based solely on these exploratory observations - more data and/or further analysis will be needed to prove their existence.

Penetration vs Speed

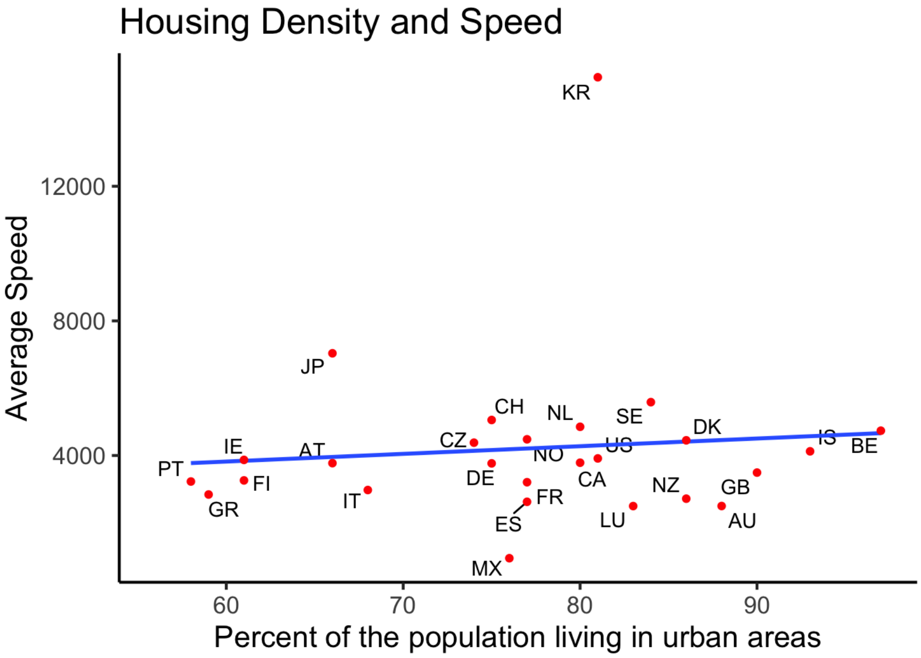
So now let’s look at the relationship between penetration and speed.

```
col_name_y = "Average.actual.speed..Akamai..kbps."
col_name_x = "Penetration.per.100.OECD..2008"

ggplot(Full_Data, aes(Penetration.per.100.OECD..2008, Average.actual.speed..Akamai..kbps., label = Full_Data[[2
]])) +
  geom_text_repel() +
  geom_smooth(method = "lm", se = FALSE) +
  geom_point(color = 'red') +
  theme_classic(base_size = 16) +
  xlab("Penetration per 100") +
  ylab("Average Speed") +
  ggtitle("Penetration and Speed")
```



We see positive relationships between peptration and speed, such that where penetration is greater, speed is faster. This runs counter to the argument that regulations requiring increased penetration lead to reduced incentives to invest in infrastructure, and therefore lower speed. Therefore this suggests that regulations requiring increased penetration do not inherently lead to a reduction in speed. However, this could just be due to the fact that the countries with increased penetration simply have higher housing density, and therefore it is easier to acheive higher penetration rates.

```
col_name_y = "Average.actual.speed..Akamai..kbps."
col_name_x = "Percent.of.population.in.urban.areas"

ggplot(Full_Data, aes(Percent.of.population.in.urban.areas, Average.actual.speed..Akamai..kbps., label = Full_Dat
a[[2]])) +
  geom_text_repel() +
  geom_smooth(method = "lm", se = FALSE) +
  geom_point(color = 'red') +
  theme_classic(base_size = 16) +
  xlab("Percent of the population living in urban areas") +
  ylab("Average Speed") +
  ggtitle("Housing Density and Speed")
```



However, when you look at how perect of th the population living in urban areas relates to speed, there doesn't seem to be a strong relationship. We can see that most countries with higher housing density have similar broadband speeds as countries with lower housing density. This suggests that the positive relationship between penetration and speed is not due to percent of the population living in urban areas.

Speed vs Price

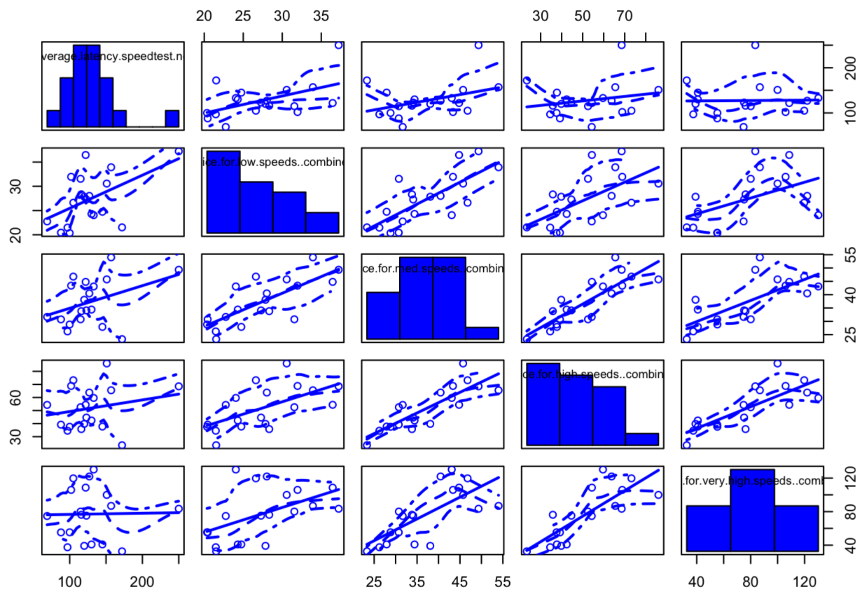
Let’s start by looking at a correlation matrix between Price variables and Speed variables. This will give us some indication regarding the correlation among variables.

```
cor(Full_Data[seq(17,31,1)], Full_Data[seq(3,6,1)], use="pairwise.complete.obs")[,c(1,2,3,4)]
```

```
                                         Price.for.low.speeds..combined Price.for.med.speeds..combined Price.for.high.speeds..combined Price.for.very.high.speeds..combined
Maximum.advertised.speed.OECD..kbps.                         -0.2020249                     -0.3039874                    -0.2211701653                          -0.43940452
Average.advertised.speed.OECD..kbps.                         -0.2131577                     -0.3602329                    -0.2929427530                          -0.61520072
Average.actual.speed..Akamai..kbps.                          -0.2657776                     -0.3018799                    -0.1357404298                          -0.38825698
Average.download.speedtest.net..kbps.                        -0.3905059                     -0.5323318                    -0.3238708804                          -0.68030255
Standard.deviation.download..speedtest.net                   -0.4010158                     -0.5043701                    -0.3159120682                          -0.78058887
Average.upload.speedtest.net..kbps.                          -0.3332764                     -0.4114289                    -0.2089438350                          -0.48674607
Standard.deviation.upload..speedtest.net                     -0.3548246                     -0.3944497                    -0.1803971783                          -0.47346969
Average.latency.speedtest.net                                 0.2861909                      0.2370186                    -0.1437633436                           0.01705699
Standard.deviation.latency..speedtest.net                     0.2679043                      0.1951490                     0.0003664236                           0.07693934
Median.download..speedtest.net..kbps.                        -0.4421358                     -0.6092223                    -0.3637093154                          -0.57735827
Median.upload..speedtest.net..kbps.                          -0.4568687                     -0.6163696                    -0.3101928820                          -0.51067237
Median.latency..speedtest.net                                 0.4752498                      0.3700521                     0.0873995886                           0.08972227
X90p..Download..speedtest.net..kbps.                         -0.3148955                     -0.4301215                    -0.2598450980                          -0.59660532
X90p..Upload..speedtest.net..kbps.                           -0.2925332                     -0.3793831                    -0.2178989124                          -0.46226157
X10p..Latency..speedtest.net                                  0.4794388                      0.3155915                    -0.0026940080                           0.10340855
```

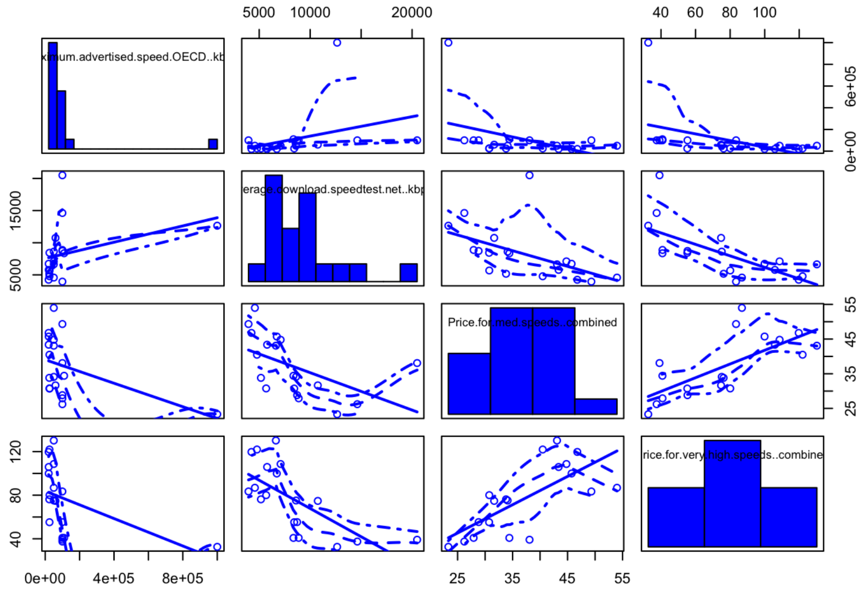
Based on the above, the average latency is a variable of interest. Comparing it agains the prices, we see that the relationship for low speed prices vs latency is a positive relationship and as the prices increase, the relationship is less strongly correlated. The relationship between prices for very high speeds vs latency is very weak.

```
scatterplotMatrix(~ Average.latency.speedtest.net + Price.for.low.speeds..combined + Price.for.med.speeds..combin
ed + Price.for.high.speeds..combined + Price.for.very.high.speeds..combined, data = Full_Data, diagonal=list(meth
od="histogram"))
```



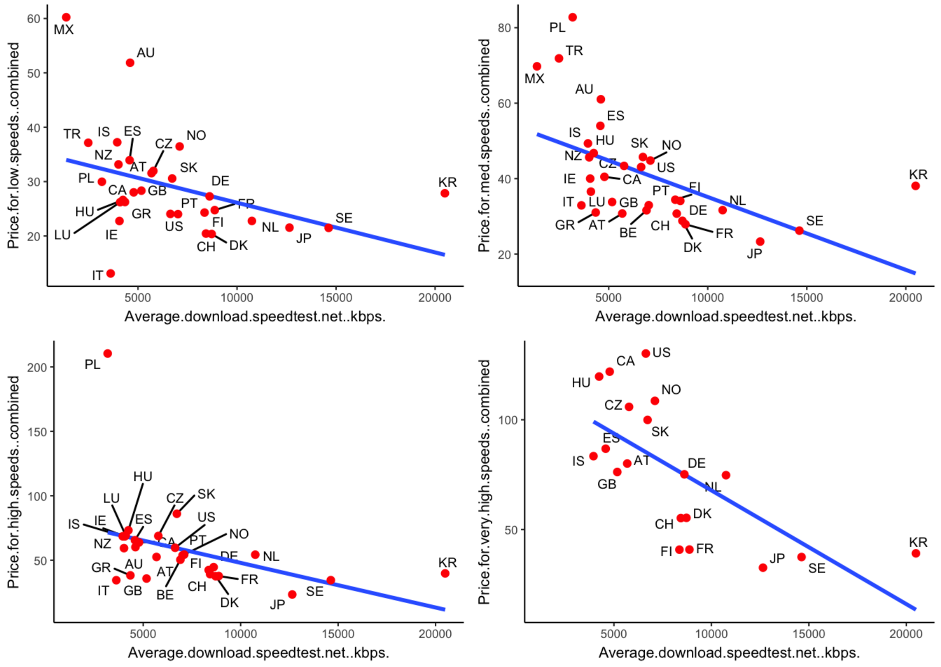
Additionally, we are interested in looking at how prices for medium and high speeds relate to maximum advertised speed and average download speeds specifically.

```
scatterplotMatrix(~ Maximum.advertised.speed.OECD..kbps. + Average.download.speedtest.net..kbps. + Price.for.med.
speeds..combined + Price.for.very.high.speeds..combined, data = Full_Data, diagonal=list(method="histogram"))
```



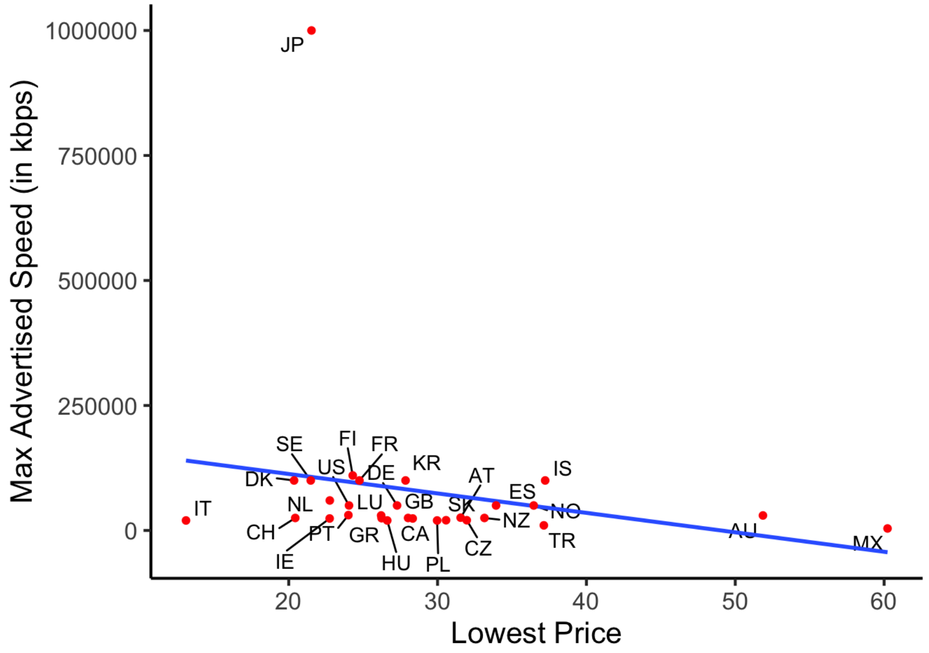
From the above scatterplot matrix, we observe that, both the Maximum Advertised Speed and Average Download speed have a negative correlation with the Price variable. Further, looking at the Maximum Advertised Speed histogram, you can see that most of the values are concentrated towards one end for the along with a few outliers.

```
p1 = ggplot(Full_Data, aes(Average.download.speedtest.net..kbps., Price.for.low.speeds..combined, label = Full_Da
ta[[2]])) +
  geom_text_repel(size = 2.5) +
  geom_smooth(method = "lm", se = FALSE) +
  geom_point(color = 'red') +
  theme_classic(base_size = 8)

p2 = ggplot(Full_Data, aes(Average.download.speedtest.net..kbps., Price.for.med.speeds..combined, label = Full_Da
ta[[2]])) +
  geom_text_repel(size = 2.5) +
  geom_smooth(method = "lm", se = FALSE) +
  geom_point(color = 'red') +
  theme_classic(base_size = 8)

p3 = ggplot(Full_Data, aes(Average.download.speedtest.net..kbps., Price.for.high.speeds..combined, label = Full_D
ata[[2]])) +
  geom_text_repel(size = 2.5) +
  geom_smooth(method = "lm", se = FALSE) +
  geom_point(color = 'red') +
  theme_classic(base_size = 8)

p4 = ggplot(Full_Data, aes(Average.download.speedtest.net..kbps., Price.for.very.high.speeds..combined, label = F
ull_Data[[2]])) +
  geom_text_repel(size = 2.5) +
  geom_smooth(method = "lm", se = FALSE) +
  geom_point(color = 'red') +
  theme_classic(base_size = 8)

plot_grid(p1, p2, p3, p4, nrow = 2, ncol = 2)
```



Comparing the average download speed measured by speedtest with the prices for varying speeds, we observe from the above plots that
- Mexico, an open access country has very high prices for low speeds and medium speeds. There is no data for high and very high speed.
- Korea on the other hand, has almost the same prices for low, medium, high and very high speeds which means that one can get access to very high speeds at not a high cost.
- For high and very high speeds, US seems to be really expensive in comparison to others.

```
ggplot(Full_Data, aes(Price.for.low.speeds..combined, Maximum.advertised.speed.OECD..kbps., label = Full_Data[[2
]])) +
  geom_text_repel() +
  geom_smooth(method = "lm", se = FALSE) +
  geom_point(color = 'red') +
  theme_classic(base_size = 16) +
  xlab("Lowest Price") +
  ylab("Max Advertised Speed (in kbps)")
```



This graph shows the highest advertised speed as measured by the OECD versus the price for low speeds. In a way, price for low speeds is the lowest advertised price since all the other prices for higher speeds are higher.

For the country of Japan, it can be easily observed that it is providing a high speed service at a low price point. This country is at the front in terms of fiber connectivity and broadband speeds.

Analysis of Secondary Effects

1. One of the secondary variables that we need to consider, which we do not have access to, is the threshold values for speed that determined each of the pricing tiers for the 4 pricing variables. As mentioned in the analysis involving Price variables, not knowing how speed factors into the determination of Low, Medium, High and Very high pricing categories causes us to question the strength and accuracy of relationships measurements between Penetration vs Price and Speed vs Price. For instance, if the speed ranges that determine a Pricing Tier is different from one country to another, or even worse overlap with another Pricing Tier in a different country, then we would need to apply some sort of transformation to speed variables prior to examining any bivariate relationships. Furthermore, lack of access to the percentage of population that subscribe to each of the Pricing Tiers add a level of difficulty in confirming some of the observed potential relationships between Price and Penetration.
2. Another confounding variable could be country size. It may be easier for smaller countries to invest in the necessary infrastructure to achieve high quality broadband with high rates of penetration. Additionally, as discussed previously, the geography of a country can also come in to play. For example, large countries with a lot of rural or mountainous land may experience additional difficulties in developing their broadband infrastructure.
3. An external variable that we don't have access to is the strength of the semiconductor industry in various countries. Countries where the semiconductor industry is in boom are more likely to invest in better infrastructure and may even have the additional funds to invest in research of cutting edge materials that increases the performance efficiency of internet technology. Having this information could provide additional insights to relationships that were identified in the previous sections, as well as information on the reasons being certain countries' leading, trailing, or outlier metrics in Speed, Penetration and Price.
4. Finally, another secondary variable could be the current state of infrastructure. Having a good infrastructure in place means that there is a higher chance of introducing state-of-the-art technologies and being up to date with the latest tools available. Countries not having a good infrastructure in place will not have the means to install the latest technologies in broadband, say for example Gigabit Fiber or availability of 4G / LTE services. As mentioned before in the price analysis, 11 out of the 30 countries did not have pricing for very high speeds which could be due to lack of a good infrastructure in place. Additionally, it was previously analyzed that the density of WiFi hotspot locations amongst the number of people is a measure of the strength of a country's WiFi infrastructure.
5. Another secondary variable of interest that we don't have access to is the knowledge about competitors in every country. Having that knowledge could affect the relationships identified before. For example, it might become clear that countries that have high level of competition among companies are offering higher broadband speeds at a comparatively low price point.

Conclusion

The 4 Price variables tiered by Speed (Low, Medium, High, Very High) are negatively correlated with 6 main Penetration variables - the 3 variables of "Penetration per 100" (OECD 2008, OECD 2007, and GC), "Household Penetration OECD", "3G Penetration per 100", and "WiFi hotspots per 100,000 (JiWire)". 5 of the penetration variables measure a percentage of the population that has access to internet service, which the 6th variable, "WiFi hotspots per 100,000" measures the strength of a country's WiFi infrastructure. We expect Price to be negatively correlated with Penetration for the 5 penetration variables that measure population access, since more consumers may choose to opt out of higher priced service. The negative correlation between Price and "WiFi hotspots per 100,000" is more difficult to explain would would require access to external variables, such as the infrastructure costs per country.

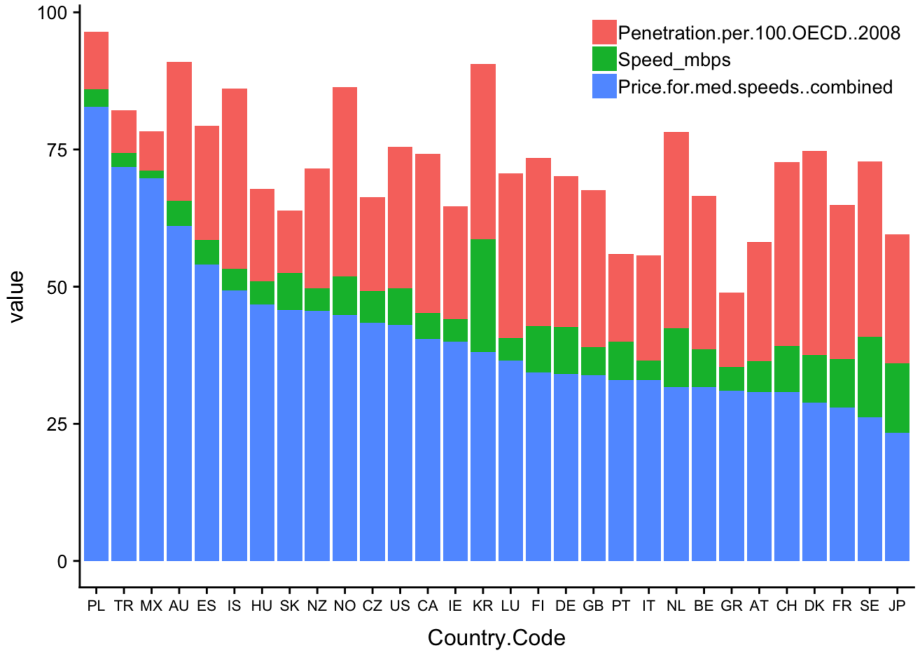
An additional conclusion from the univariate analysis on Price per Country is that each of the 3 countries which did not adopt an open access policies (Mexico, the Slovak Republic, and the United States) have one of the highest prices in at least one of the Pricing Tiers by Speed. Mexico has the highest price for Low Speeds, one of the highest prices for Medium Speeds; Slovak Republic has one of the highest prices for High Speeds; The United States has the Highest Price for Very High Speeds.

Speed is negatively correlated with price, such that countries where average higher speeds measured have lower prices. This is counter to what is expected if countries have a tiered pricing system where price for faster services is higher. It could mean that the either these countries have a really good infrastructure set up that they can offer such high speeds at a low price point or that there is an increased level of competition among companies in these countries that they are willing to offer high speeds at a low price.

Penetration is positively correlated with speed, such that countries with higher rates of penetration have faster broadband connections. This runs counter to the argument that regulations requiring increased penetration reduce incentives to improving infrastructure. If that argument were true, we would expect that increased penetration would be correlated with decreased speed, but instead, we see the opposite relationship.

One way to view the all three variables simulteneously can be seen below. For each of the concepts, Speed, Price and Penetration, we select one representative variable, based on univariate analysis, to plot in the stacked bar chart below by Country. Please note that the representative Speed Variable has been changed to mbps units

```
Full_Data$Speed_mbps = Full_Data$Average.download.speedtest.net..kbps./1000

Full_Data$Country.Code <- factor(Full_Data$Country.Code, levels = Full_Data$Country.Code[order(-Full_Data$Price.for.med.speeds..combined)])
Full_Data_tf <- melt(Full_Data[, c("Country.Code", "Penetration.per.100.OECD..2008", "Speed_mbps", "Price.for.med.speeds..combined")], id.vars = 1)

Full_Data_tf %>%
  ggplot(aes(x = Country.Code, y = value)) +
  geom_bar(stat = "identity", aes(fill = variable), position = "stack") +
  theme(legend.justification=c(1,1),legend.position=c(1,1),legend.title=element_blank(), text=element_text(size=12), axis.text.x=element_text(size=8), axis.text.y=element_text(size=10))
```



Based on the above observations, we do not see evidence of trade offs between price, penetration, and speed. In fact, we see several countries, notably Japan and South Korea, that excel in all three. On the flip side, as demonstrated by countries such as Poland, having open access policies is clearly not sufficient to having high quality, accessible broadband. It is likely that some of the secondary variables mentioned above have a substantial effect on the quality and accessibility of broadband in a country. However, due to the lack of evidence of trade offs between price, penetration, and speed, assertions that open access policies would lead to such trade offs are unwarranted. Since there are only three countries in this dataset without open access policies, it is hard to draw clear conclusions about the benefits of open access policies compared to free markets. But based on the evidence at hand, the arguments against open access policies appear to be erroneous.