Broadband Exploratory Data Analysis (Lab #1)

w203.1 Team (Satya Thiruvallur,Sudha Subramanian & Chase Inguva)

Introduction

This analysis is motivated by the following research questions:

- RQ1: Is there a three-way relationship between price, speed, and penetration?
- RQ2: Does a trade off exist between price, speed and penetration?
- RQ3: Is there evidence for the beneficial effects of open access policies?

We will address this question using exploratory data analysis techniques. Our data comes Harvard Berkman Center study on next generation connectivity. The data was compiled from multiple sources and using subject matter experts in this area.

Data Overview

We are provided price, penetration and speed data for 30 countries. The penetration data has a reference to years 2007 and 2008 for OECD penetration metrics. Few Caveats:

1. We are uncertain about the timing of the data sets as there is no reference to 4G connectivity. We assume the price, speed and penetration data are of the same time period, potentially 2008 or 2009.
2. We assume data in price, penetration and speed to be numerical variables other than country name and country code.
3. The current analysis is limited to descriptive statistics.
4. Our analysis is not causal, however we look at correlations of variables.

Setup

As a first step in our analysis, we load the required libraries as captured below. These libraries allow us to perform the tranformations and exploratory data analysis.

```
require(dplyr)
require(ggplot2)
require(stringr)
require(plotly)
require(gridExtra)
require(tidyr)
require(corrplot)
require(data.table)
```

We utilize the Rmd file for our analysis and the rmd file contains code markup as well as the analysis. We save the data files in in the same directory as our Rmd file.

Loading Data Sets

We load price, penetration and speed csv files into data frames.

```
# Load the data from csv
price = read.csv("Price.csv", stringsAsFactors = FALSE)
penetration = read.csv("Penetration.csv", stringsAsFactors = FALSE)
speed = read.csv("Speed.csv", stringsAsFactors = FALSE)
```

Variables Overview

The datasets used in this exploratory analysis include:

- Price: Dataset contains categorical variables, country name and two character country code and numerical variables describing price information for low, med, high and very high speed internet packages across 30 countries. We are assuming the prices are for retail consumers. There are
- Penetration: Dataset contains categorical variables, country name and two character country code and numerical variables describing penetration information across various factors across 30 countries. One row in the dataset had no recorded country / country code and values in other columns are NAs.
- Speed: Dataset contains categorical variables, country name and two character country code and numerical variables to measure speed of internet packages. The key items reported are advertized, download, upload speeds and latency and across 30 countries. This file also contained two rows with blank or NAs across all columns.

Data Preperation

Our data preperation process included multiple steps such as merging three data sets, renaming columns, handling data quality issues and data re-shaping. Each of these items are described in further detail in the next sections.

Merging Datasets

We merged price, penetration and speed data frames using an inner join function by joining on country code column that is the key in each of the datasets being merged. The country code column was in mixed case in price and penetration data frame and in lowe case in speed data frame. We had to account for the column name inconsistency and as there are three datasets, this is done as a 2-step process. Due to the inner join, the rows with NA's in penetration and speed file were removed. The output data frame is all_comb.

```
price_penet <- inner_join(price, penetration, by=c("Country.Code", "Country"))
all_comb <- inner_join(price_penet, speed, by=c("Country.Code" = "Country.code", "Country" = "Country"))
```

Rename Columns

The raw data had very long column names as well as special characters. We renamed the columns names using 'rename' function in dplyr package to simplify the analysis.

Data Clean-up

We ran summary statistics on the merged data set(see appendix) and identified the data cleanup items. The merged data frame had inconsistent formats for specific columns although they appear like numbers. Price, Percentages and Speed metrics were cleaned up where the value is a character variable and was converted to numeric values. Key issues encountered during this cleanup was removing punctuation characters, $, % etc. The code section below captures how this was accomplished.

```
# some cleanup of factor variables (price) --> converted to numeric
all_comb$PriceHighSpeed <- as.numeric(gsub("\\$", "", all_comb$PriceHighSpeed))
all_comb$PriceLowSpeed <- as.numeric(gsub("\\$", "", all_comb$PriceLowSpeed))
all_comb$PriceMedSpeed <- as.numeric(gsub("\\$", "", all_comb$PriceMedSpeed))
all_comb$PriceVeryHighSpeed <- as.numeric(gsub("\\$", "", all_comb$PriceVeryHighSpeed))
# convert to numeric type for the purpose of analysis (remove %)
all_comb$Growth.in.3G.penetration <- as.numeric(gsub("\\%", "", all_comb$Growth.in.3G.penetration))
all_comb$Pop_Urban <- as.numeric(gsub("\\%", "", all_comb$Pop_Urban))
# converting some character columns to numeric after removing the commas which introduce NAs if convert
all_comb$Advt_MaxSpeed <- as.numeric(gsub("\\,","", all_comb$Advt_MaxSpeed))
all_comb$Advt_AvgSpeed <- as.numeric(gsub("\\,", "", all_comb$Advt_AvgSpeed))
all_comb$Actual_AvgSpeed <- as.numeric(gsub("\\,", "", all_comb$Actual_AvgSpeed))
all_comb$Download_AvgSpeed <- as.numeric(gsub("\\,", "", all_comb$Download_AvgSpeed))
all_comb$Download_STD <- as.numeric(gsub("\\,", "", all_comb$Download_STD))
all_comb$Upload_AvgSpeed <- as.numeric(gsub("\\,", "", all_comb$Upload_AvgSpeed))
all_comb$Upload_STD <- as.numeric(gsub("\\,", "", all_comb$Upload_STD))
all_comb$Latency_Avg <- as.numeric(gsub("\\,", "", all_comb$Latency_Avg))
all_comb$Latency_STD <- as.numeric(gsub("\\,", "", all_comb$Latency_STD))
all_comb$Download_Median <- as.numeric(gsub("\\,", "", all_comb$Download_Median))
all_comb$Upload_Median <- as.numeric(gsub("\\,", "", all_comb$Upload_Median))
all_comb$Download_X90p <- as.numeric(gsub("\\,", "", all_comb$Download_X90p))
all_comb$Upload_X90p <- as.numeric(gsub("\\,", "", all_comb$Upload_X90p))
all_comb$X <- NULL
```

Reshaping Data for Analysis

all_comb data frame has columns which indicate price for speeds and we tranformed the price variables using dplyr's 'gather' function to form columns - one for category and another for value that is derived from a set of columns. The purpose of this was to perform visualizations in a single plot across the different categories.

```
price_comb <- all_comb %>%
  gather(Speed, CombPrice, PriceLowSpeed:PriceVeryHighSpeed) %>%
  select(Country, Country.Code, Speed, CombPrice, Penet_2008, Penet_2007)

penet_comb <- all_comb %>%
  gather(PenetrationCategory, Penetration, Penet_2008:Penet_3G) %>%
  select(Country, Country.Code, PenetrationCategory, Penetration, Growth.in.3G.penetration, Pop_Urban)

speed_comb <- all_comb %>%
  gather(SpeedCategory, Speed, Advt_MaxSpeed:Latency_Median) %>%
  select(Country, Country.Code, Pop_Urban, SpeedCategory, Speed)
```

Exploratory Data Analysis

After data preperation, we explore the data to understand the descriptive statistics. We perform univariate analysis and bivariate analysis to understand the price - speed - penetration relationship. The following sections provide overview of univariate analysis, bivariate analysis, potential secondary effects and conclusions.

Univariate Analysis

We analyze price, penetration and speed variables individually and explain interesting behavior. We start our analysis with price for internet packages.

Price

We begin our analysis looking at countries with missing values for the four pricing options - low, medium, high and very high. We observe the following missing values:

- Very high speed: 11 countries are missing price information (NAs) for "Very High Speed"
- High speed: Mexico & Turkey are missing price information (NAs) for "High Speed"
- Medium speed: All countries have pricing data
- Low speed: Belgium has missing price for "Low Speed", however it has data for medium and high speeds.

We utilize a histogram to check the price variation across speed levels and countries. We observe few outliers for low speeds and high speeds in our data. In order to understand these plots side-by-side, they are arranged in a 2x2 grid using par and grid.arrange function calls. In the code sequence below, a function 'myqplot' is defined that takes in the column name, label to use and binwidth as parameters to render the plot. The plots for each category are built by invoking this fucntion.

```
myqplot <- function(colname, collabel, usebinwidh) {
  qplot(colname, geom="histogram", main=paste0("Histogram - Price for", collabel),
        xlab=paste0("Price-", collabel), ylab="Count",
        fill=I("blue"), col=I("red"), alpha=I(.2), binwidth = usebinwidh)}
plow <- myqplot(all_comb$PriceLowSpeed, "Low Speeds", 3)
pmed <- myqplot(all_comb$PriceMedSpeed, "Med Speeds", 3)
phigh <- myqplot(all_comb$PriceHighSpeed, "High Speeds", 10)
pveryhigh <- myqplot(all_comb$PriceVeryHighSpeed, "Very High Speeds", 5)
par(mfrow = c(2,2))  # 2 rows and 2 columns
grid.arrange(plow, pmed, phigh, pveryhigh)
```

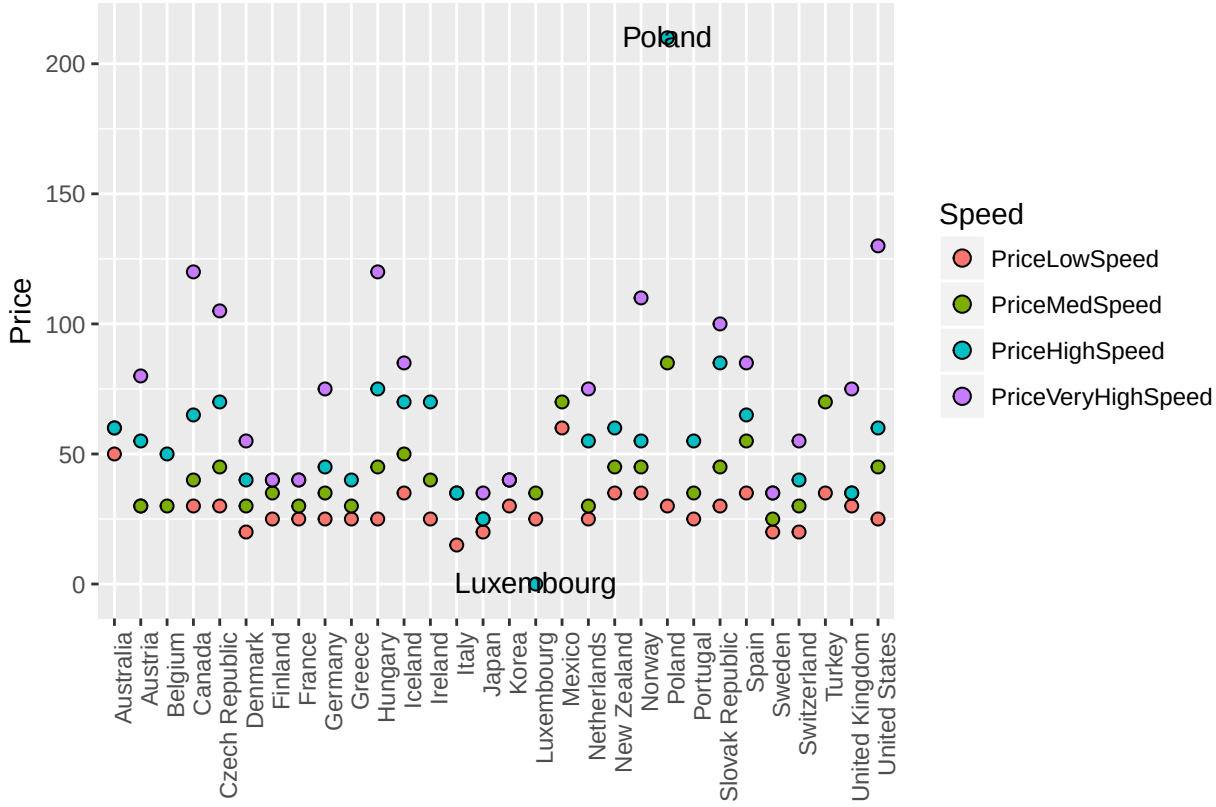
As can be seen above, there is some overlap in the ranges of x-axis across the different categories, which is

expected. However, it is clear that there is an outlier (High Speed category, which has a value much higher than the average in that category). While the plots for low, med and high speeds are unimodal, very high speed category depicts a bimodal plot.

Price Anomalies

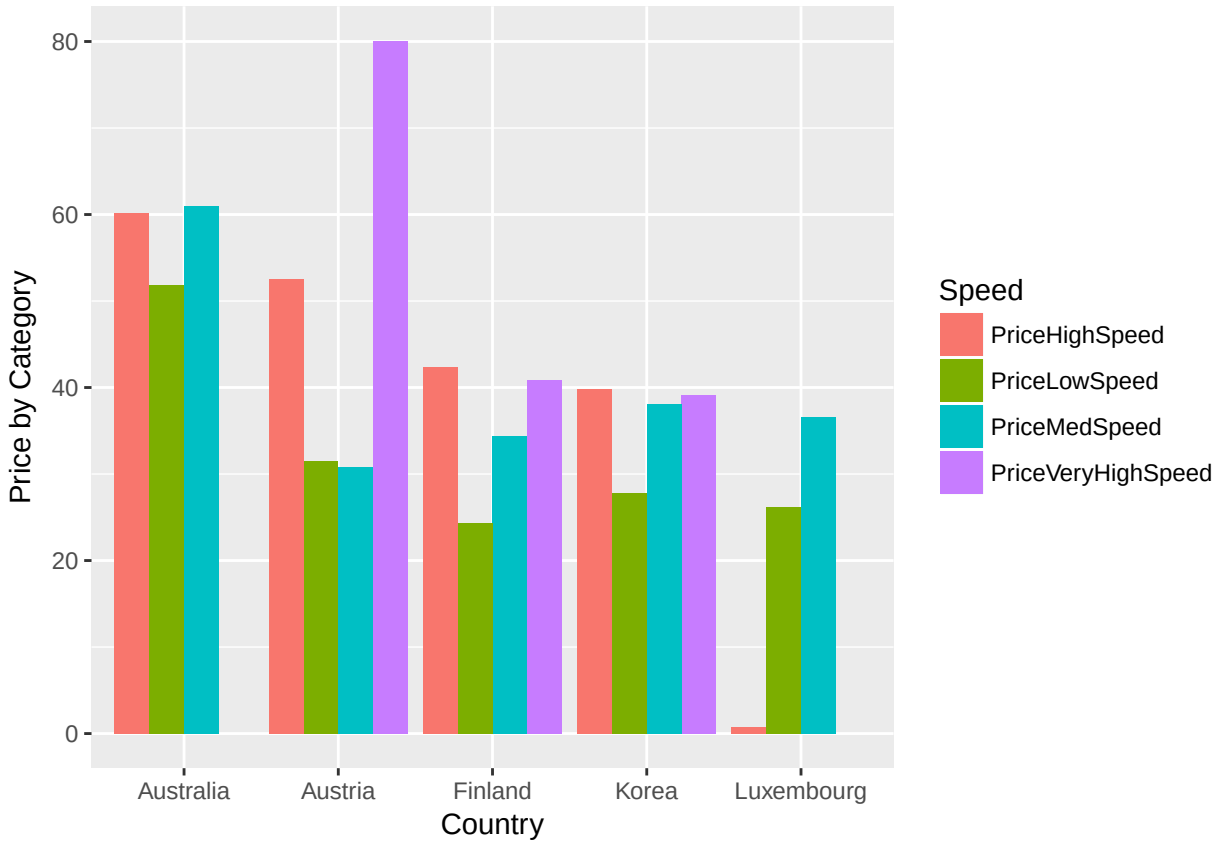
We plot a dot plot with the four price variables to identify any anomalies. Poland and Luxemburg stand out with the price anomalies for high speed package as shown below.

```
price_comb$Speed <- factor(price_comb$Speed, levels = c("PriceLowSpeed", "PriceMedSpeed",
                                                        "PriceHighSpeed", "PriceVeryHighSpeed"))
par(mfrow = c(2,2))
price_comb$name <- rownames(price_comb)
ggplot(price_comb, aes(x=Country, fill=Speed, y=CombPrice)) +
  geom_dotplot(binaxis="y", method = "histodot", binwidth = 5, stackdir = "center") +
  geom_text(data=subset(price_comb, (CombPrice > 200) | (CombPrice < 1)) ,aes(label=Country)) +
  theme(axis.text.x = element_text(angle=90, hjust=1)) + xlab("") + ylab("Price")
```



To further isolate further outliers in the data set, we look at scenarios where price across tiers (low, med, high and very high categories) is inconsistent. We identify the following anomalies:

```
# plot that captures info where price of lower tier is higher than that of the higher tier
# Luxemourg clearly has the data wrong - price of very high is way less than low / medium
all_comb %>%
  filter((PriceVeryHighSpeed < PriceHighSpeed) | (PriceHighSpeed < PriceMedSpeed) | (PriceMedSpeed < Pr
  gather(Speed, CombPrice, PriceLowSpeed:PriceVeryHighSpeed) %>%
  ggplot(aes(x=Country, y=CombPrice)) +
  geom_bar(stat="identity", position="dodge", aes(fill=Speed)) +
  labs(x="Country", y="Price by Category")
```



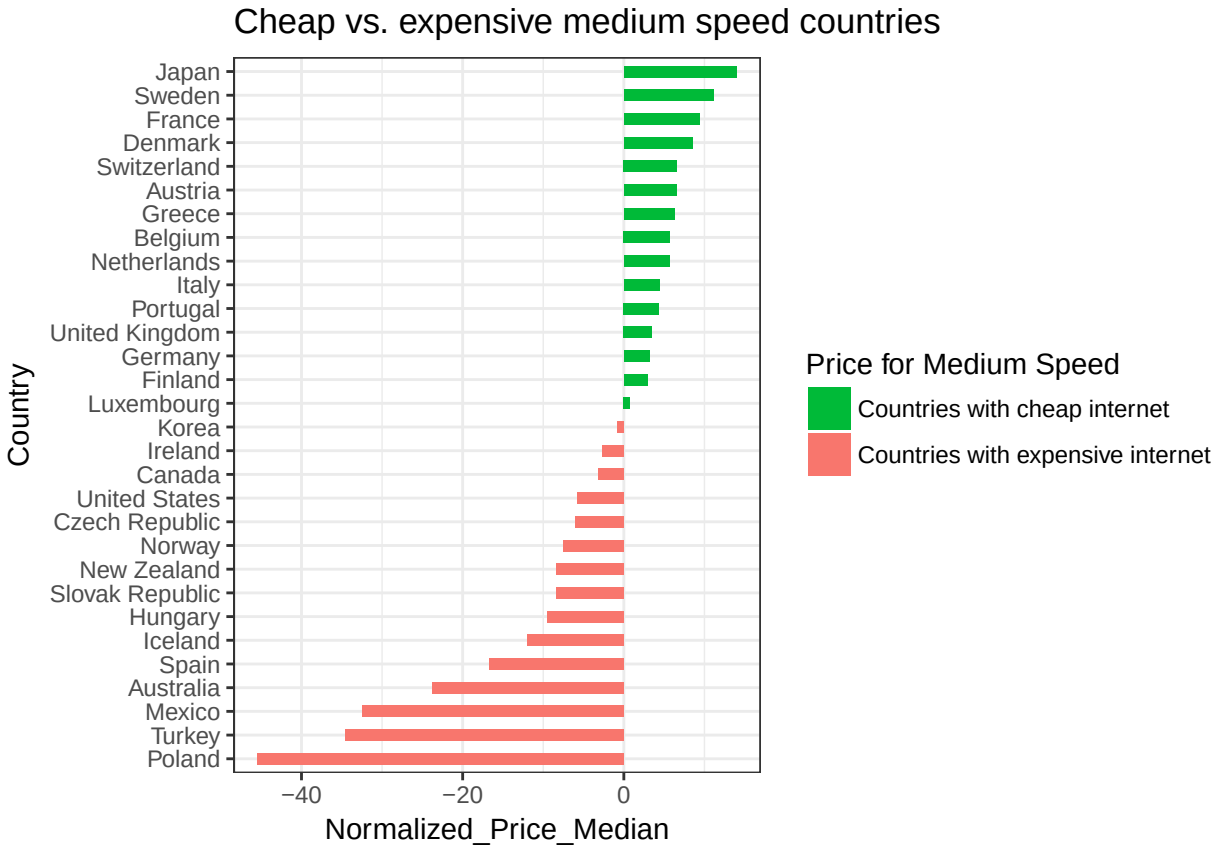
Price comparision across countries

We use price medium speed to compare countries as it doesn't have any missing values or large outliers. We develop a diverging bar chart to identify countries with cheapest medium speed internet by calculating:

- Normalized_Price_Median = -1 * (Median of PriceMedSpeed - PriceMedSpeed of country)

Japan, Sweden, France are the countries with the cheapest medium speed internet vs. Poland, Turkey and Mexico have the most expensive mediun speed internet.

```
theme_set(theme_bw())
# Data Prep
#First subset the dataframe to the the dataframe of interest
med_speeds_study <- subset(all_comb, select = c("Country", "PriceMedSpeed"))
med_speeds_study$Normalized_Price_Median <- round(med_speeds_study$PriceMedSpeed -
                                                    median(med_speeds_study$PriceMedSpeed), 2) * -1
med_speeds_study$Normalized_direction <- ifelse(med_speeds_study$Normalized_Price_Median < 0 , "below",
med_speeds_study <- med_speeds_study[order(med_speeds_study$Normalized_Price_Median), ]
med_speeds_study$Country <- factor(med_speeds_study$Country, levels= med_speeds_study$Country)
#Diverging Barcharts
ggplot(med_speeds_study, aes(x=Country, y=Normalized_Price_Median, label=Normalized_Price_Median)) +
  geom_bar(stat='identity', aes(fill=Normalized_direction), width=.5) +
  scale_fill_manual(name="Price for Medium Speed",
                    labels= c("Countries with cheap internet", "Countries with expensive internet"),
                    values = c("above"="#00ba38", "below"="#f8766d")) +
  labs(title= "Cheap vs. expensive medium speed countries") +
  coord_flip()
```



Penetration

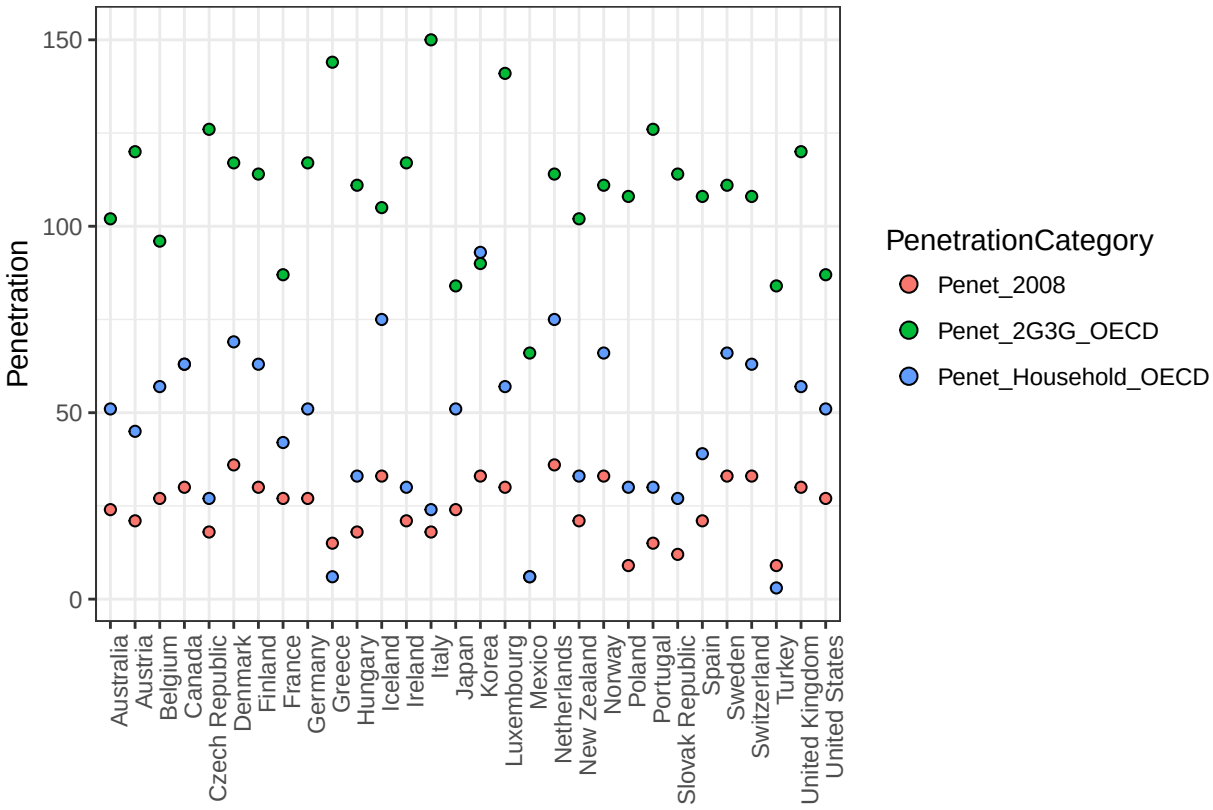
Penetration data provides summary of penetration metrics for households, Wifi hotspots, 2G & 3G penetration as well as percentage of urban population. We start our analysis by running summary stats of the penetration columns(see Appendix). Percentage of urban population column stands out with an unexpected outlier. Poland has 162% population in urban areas and it is an invalid data point.

```
#df2 <- subset(all_comb, select = c(7:16))
#summary(df2)
summary(all_comb$Pop_Urban)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.00   66.25   77.00   78.33   83.75  162.00
```

We then look at 3G penetration per 100, 2G & 3G penetration per 200 and Household penetration per 100. Looking at each variable, we see Korea has highest household penetration.

```
# use this for penetration univariate plot; Korea has highest household penetration
penet_comb %>% filter((PenetrationCategory == "Penet_2008")
                      | (PenetrationCategory == "Penet_2G3G_OECD") | (PenetrationCategory == "Penet_Hous
  ggplot(aes(x=Country, fill=PenetrationCategory, y=Penetration)) +
  geom_dotplot(binaxis="y", method = "histodot", binwidth = 3, stackdir = "center") +
  # geom_text(data=subset(penet_comb) ,aes(label=Country)) +
  theme(axis.text.x = element_text(angle=90, hjust=1)) + xlab("") + ylab("Penetration")
```

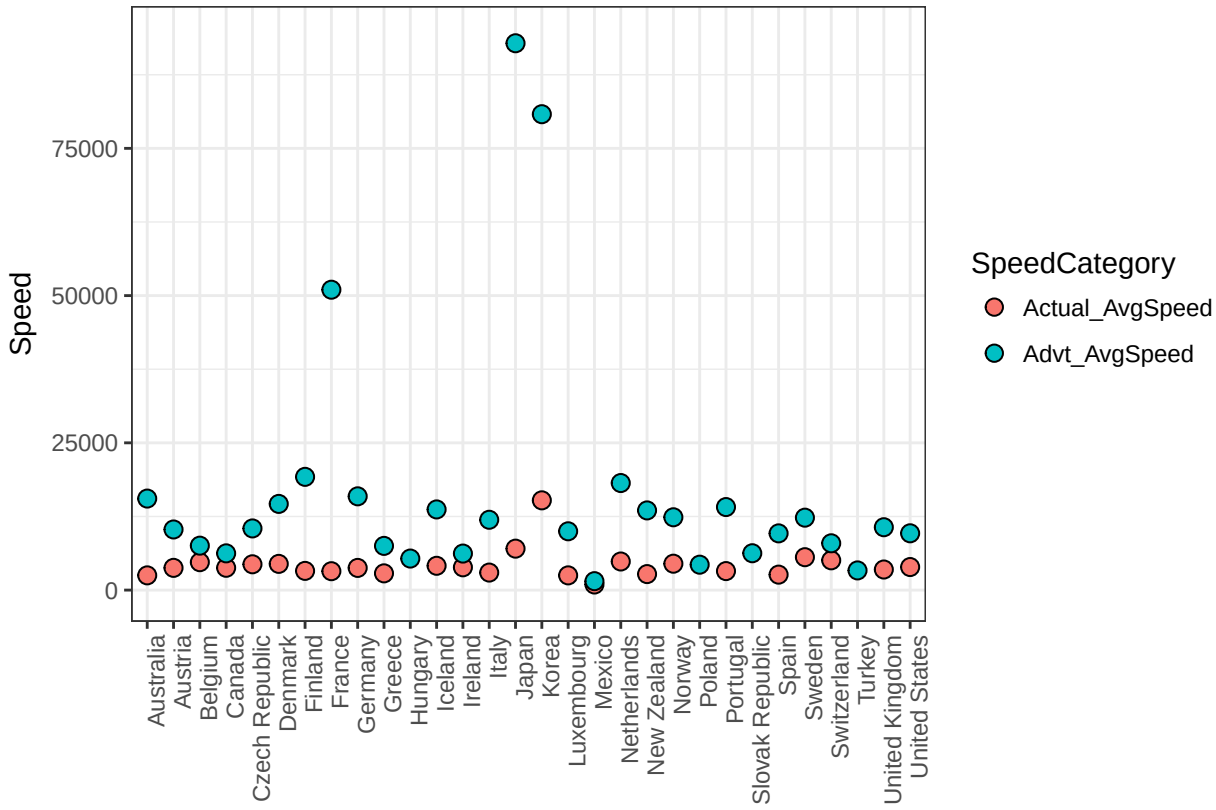


Speed

Speed data provides summary of speed metrics for countries. The metrics include download speeds, upload speeds, advertized and actual speeds. We initally run summary statistics for the data and identify download and upload speeds are most useful metrics for the users. We first analyze average advertized internet speeds compare to average actual speeds. Italy, Japan & France have the highest advertized average speeds and Korea has the highest average actual speed.

```
speed_comb %>%
  filter((SpeedCategory == "Actual_AvgSpeed") | (SpeedCategory == "Advt_AvgSpeed")) %>%
  ggplot(aes(x=Country, fill=SpeedCategory, y=Speed)) +
  geom_dotplot(binaxis="y", stackdir = "center") +
  theme(axis.text.x = element_text(angle=90, hjust=1)) + xlab("") + ylab("Speed")
```

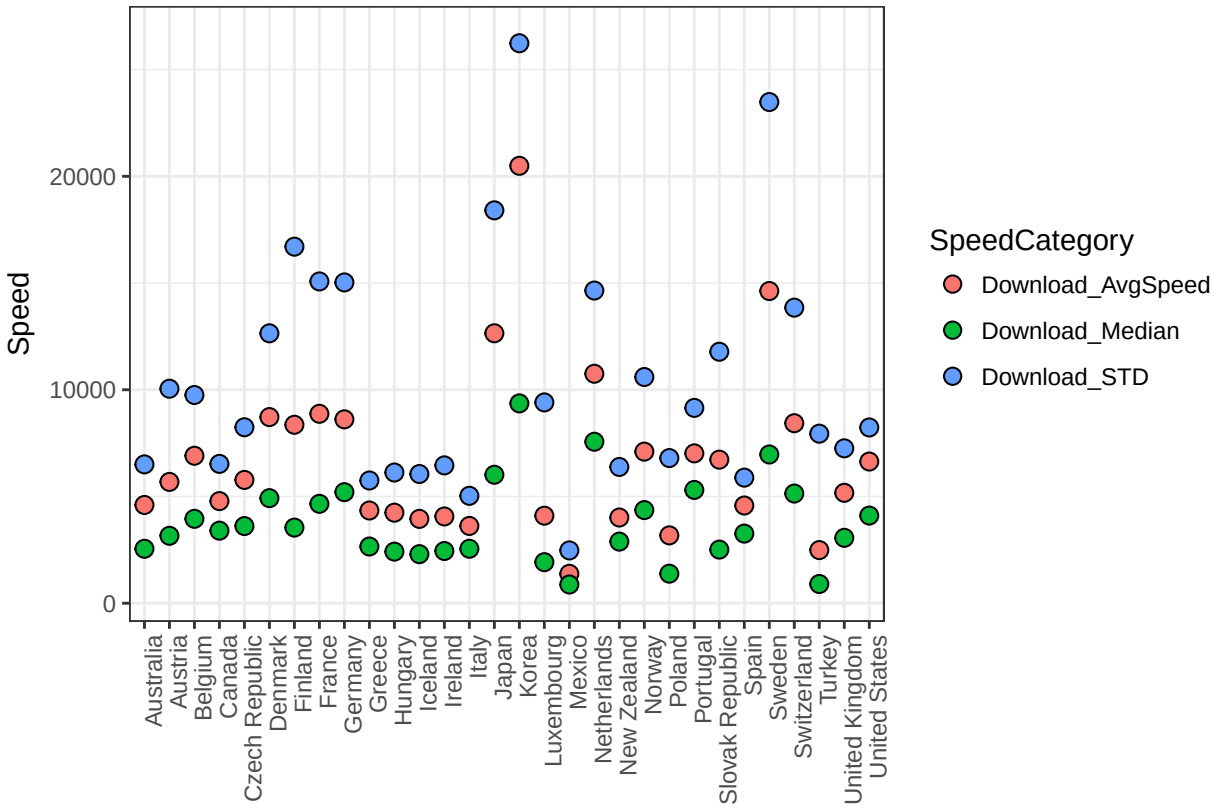
```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



We then analyze the download speed statistics. Italy & Japan have the highest average download speeds and Korea & Sweden have the highest standard deviation of average download speeds.

```
speed_comb %>%
  filter((SpeedCategory == "Download_Median") | (SpeedCategory == "Download_AvgSpeed")  |
           (SpeedCategory == "Download_STD")) %>%
  ggplot(aes(x=Country, fill=SpeedCategory, y=Speed)) +
  geom_dotplot(binaxis="y", stackdir = "center") +
  theme(axis.text.x = element_text(angle=90, hjust=1)) + xlab("") + ylab("Speed")
```

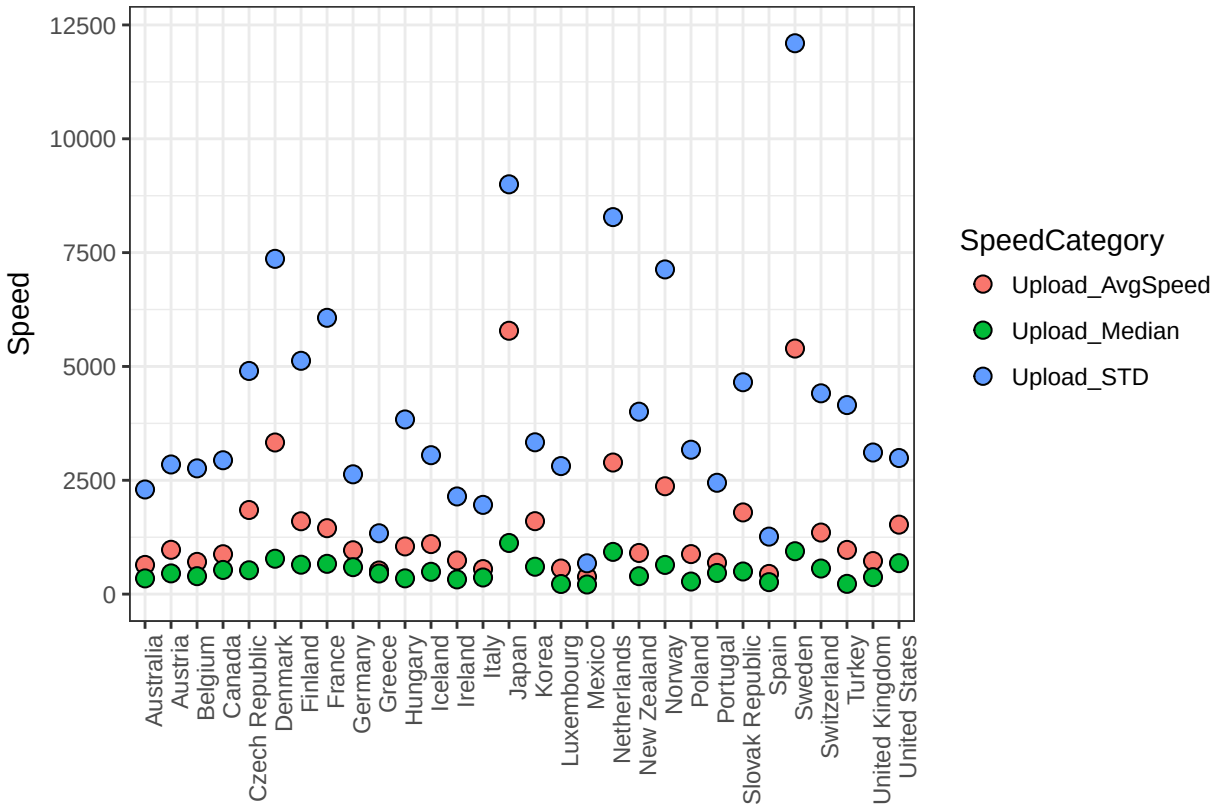
```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



We next analyze the upload speed statistics. Japan & Sweden have the highest average upload speeds and Sweden, Japan & Netherlands have the highest variability in average upload speeds.

```
speed_comb %>%
  filter((SpeedCategory == "Upload_Median") | (SpeedCategory == "Upload_AvgSpeed")  | (SpeedCategory ==
  ggplot(aes(x=Country, fill=SpeedCategory, y=Speed)) +
  geom_dotplot(binaxis="y", stackdir = "center") +
  theme(axis.text.x = element_text(angle=90, hjust=1)) + xlab("") + ylab("Speed")
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```

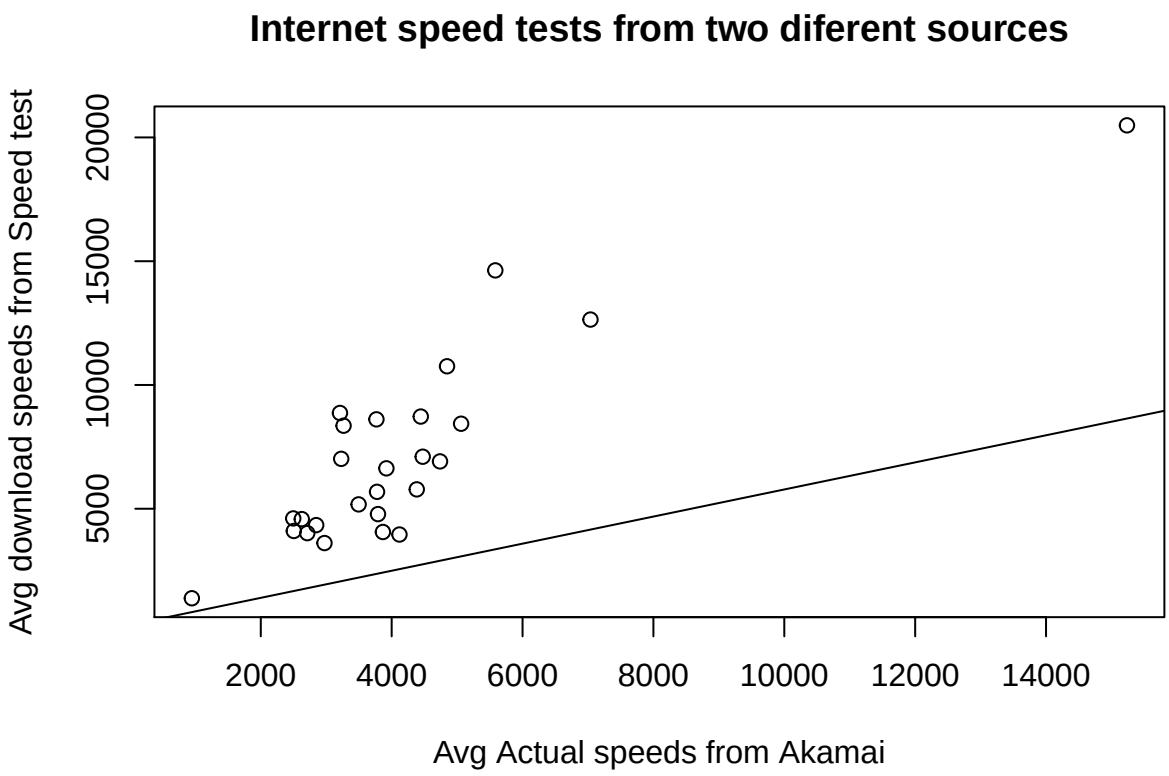


In the next section, we look at bivariate relationships between variables.

Bivariate Analysis

This scatter plot will attempt to verify the correlation between Internet speed test numbers generated from 2 different sources: Speedtests.net & Akamai

```
plot(jitter(all_comb$Actual_AvgSpeed, factor=2), jitter(all_comb$Download_AvgSpeed, factor=2),
xlab = "Avg Actual speeds from Akamai", ylab = "Avg download speeds from Speed test",
main = "Internet speed tests from two diferent sources")
abline(lm(all_comb$Actual_AvgSpeed ~ all_comb$Download_AvgSpeed))
```



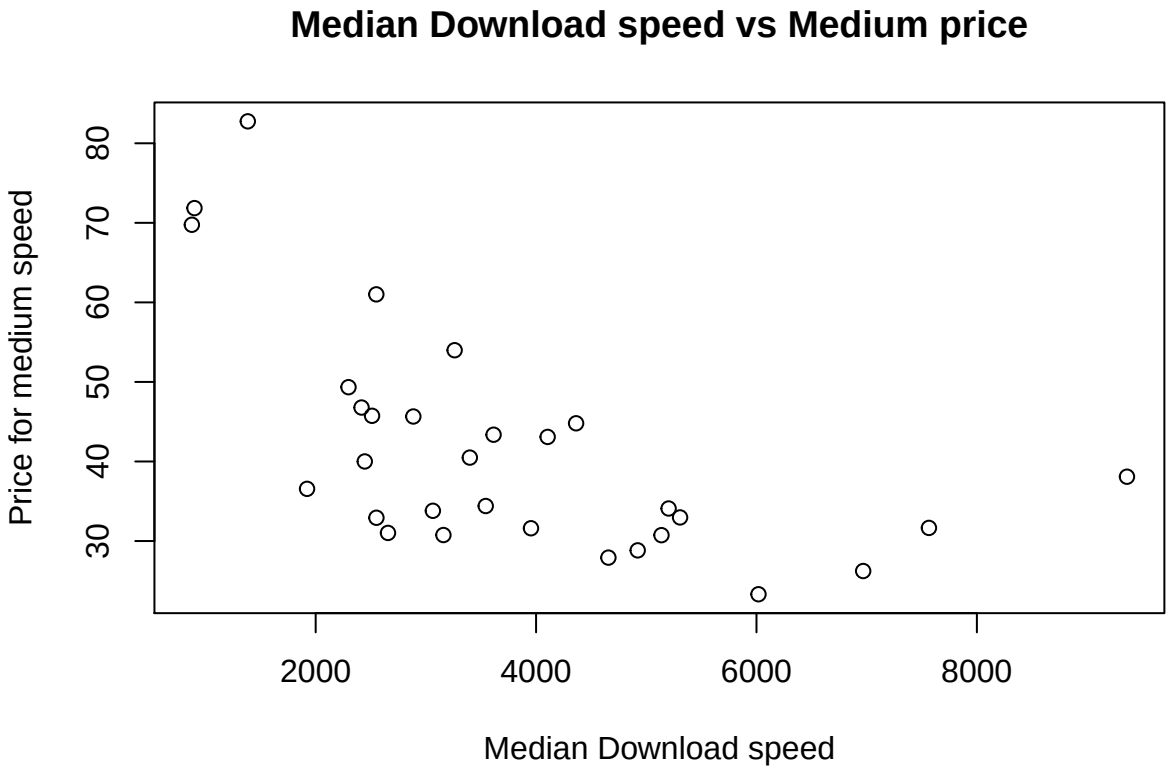

```
# We will also check the correlation score between these 2 vectors
cor(all_comb$Actual_AvgSpeed, all_comb$Download_AvgSpeed, use = "complete.obs")
```

```
## [1] 0.8679671
```

We notice that there is a strong positive correlation between the Speed numbers reported by 2 different agencies. This lends more credibility to the Speed dataset and confidence to our analysis

This scatter plot will attempt to verify the correlation between Median download speed #available in all countries vs Price for medium internet speed. The goal for this char is to #understand the degree of correlation between the most common speed and most common price for the speed.

```
# Scatter ploot between Median Download speed and Price for Medium speed
plot(jitter(all_comb$Download_Median, factor=2), jitter(all_comb$PriceMedSpeed, factor=2),
xlab = "Median Download speed", ylab = "Price for medium speed",
main = "Median Download speed vs Medium price")
abline(lm(all_comb$Download_Median ~ all_comb$PriceMedSpeed))
```



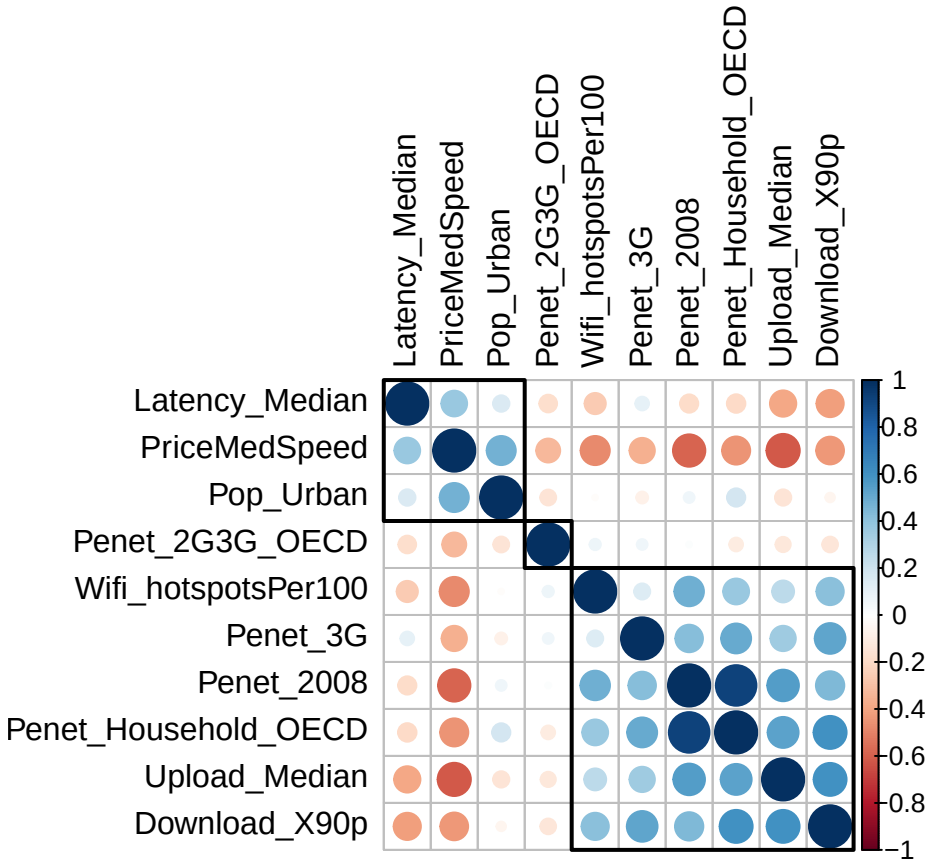

```
cor(all_comb$Download_Median, all_comb$PriceMedSpeed, use = "complete.obs")
```

```
## [1] -0.6092223
```

We see that there is a strong negative correlation between these variables. The lower the price for internet speed , higher the download speed

We utilize the most important variables from the univariate analysis of price, penetration and speed and create a correlation plot. This correlation plot helps us review the relationship between price, penetration and speed closely

```
df2 <- subset(all_comb, select = c(4,7,9,10, 12,15,16,27,28,29))
tway <- cor(df2)
corrplot(tway, order = "hclust", addrect = 3,tl.col="black")
```



We can see a strong negative correlation between Price for Medium speed vs Wifi hotspots per 100 Median Download and Upload speed Household penetration This makes sense since the lower price is indeed caused by a higher performing network infrastructure with deep penetration. For the same reason , The Download/Upload speeds have strong correlation with Household penetration and Wifi Hotspots per 100

Cumulative Rank Analysis

We attempted to define a cumulative rank combining 3 pertinent metrics:Penetration, Speed and Price. We choose the Household penetration , Price for Medium speed to represent penetration and price metrics respectively. For speed, we came up with new metric "Download_Upload" that is a scaled product of Download and Upload metrics.

The cumulative rank considered the following ratio for each: 50% for penetration 35% for speed 15% for price Penetration has been given the highest importance followed by speed and price to measure performance. We believe this ratio to reflect the priority for assessment of performance for the network industry. The dataset was ordered based on this rank and plotted as an ordered bar char here

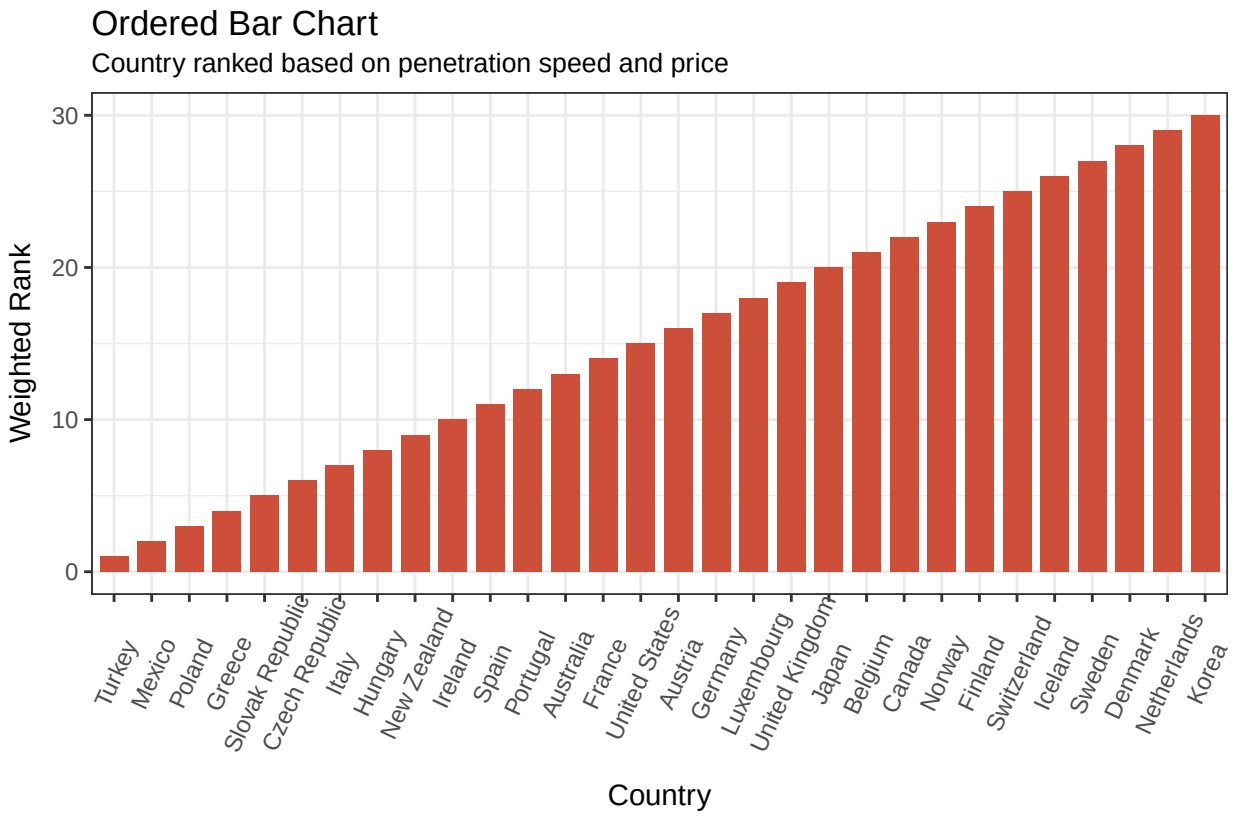
```
# weighted ranking / graph
rank_df <-  subset(all_comb, select = c("Country","Penet_Household_OECD","Download_Median","Upload_Media
rank_df$Median_download_upload <- scale(round((rank_df$Download_Median * rank_df$Upload_Median)/1000, 2
setnames(rank_df, old=c("Penet_Household_OECD","PriceMedSpeed","Median_download_upload"), new=c("Househo
rank_df$Household_Pentrn <- scale(rank_df$Household_Pentrn)
rank_df$Price_med_speed <- scale(rank_df$Price_med_speed)
rank_df$Download_Median <- NULL
rank_df$Upload_Median <- NULL
rank_df$weighted_score <- ( rank_df$Household_Pentrn * 0.50 ) + ( -rank_df$Price_med_speed * 0.15 ) + (
rank_df$rank_mvar <- rank(rank_df$weighted_score, na.last = TRUE, ties.method = "average")
```

```r
rank_df <- rank_df[order(rank_df$rank_mvar, na.last = TRUE, decreasing = FALSE),]
rank_df$Country <- factor(rank_df$Country, levels = rank_df$Country)
tail(rank_df, n = 10)
```

```
##        Country Household_Pentrn Price_med_speed Median_up_down_speed
## 3      Belgium        0.4531782     -0.69008910          -0.35038138
## 4       Canada        0.8077249     -0.06798493          -0.22367947
## 21      Norway        0.9221241      0.23430019           0.28510314
## 7      Finland        0.7478015     -0.49370898           0.02455859
## 27 Switzerland        0.7532491     -0.75180857           0.33172312
## 12     Iceland        1.3479434      0.55271653          -0.57191862
## 26      Sweden        0.9175845     -1.06812084           2.21760602
## 6      Denmark        1.0496883     -0.88576787           0.81230188
## 19 Netherlands        1.2421695     -0.68728367           2.44788451
## 16       Korea        2.1664424     -0.23631075           1.73678616
##    weighted_score rank_mvar
## 3       0.3288761        21
## 4       0.4132773        22
## 21      0.4269149        23
## 7       0.4480431        24
## 27      0.4905569        25
## 12      0.5890625        26
## 26      0.6267720        27
## 6       0.6605524        28
## 19      0.7327449        29
## 16      1.1247466        30
```

```r
# Graph Set up work
theme_set(theme_bw())

# Draw plot
ggplot(rank_df, aes(x=Country, y=rank_mvar)) +
  geom_bar(stat="identity", width=.75, fill="tomato3") +
  labs(title="Ordered Bar Chart",
       subtitle="Country ranked based on penetration speed and price",
       caption="source: broadband eda",
       y = "Weighted Rank" ) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
```



Based on the ranking we notice that Korea is the leader amongst all countries and Turkey forms the tail of this list. It is also interesting to note that United States sits in the middle based on this rating.

Analysis of Secondary Effects

We are able to see through the bi-variate analysis and plots the effects of penetration and speed on price. With the information available in the three datasets used in this analysis, we have understood how price, penetration and speeds vary across the 30 countries in the dataset. Of the 30 countries, United States, Mexico and Slovak Republic have not opted for Open Access Policy. Considering that this information is given, although not available in the dataset used for analysis, we have put together some visualizations to help understand the impact of adopting to open access policies.

For this purpose, additional columns to capture the category, as well as percent difference in price between tiers are defined as shown below.

```
# Analysis of Secondary Effects
all_comb$OpenAccess <- ifelse((all_comb$Country.Code %in% c("US", "MX", "SK")), "Non-Adopt", "Adopt")
all_comb$diffhightoveryhigh <- (all_comb$PriceVeryHighSpeed - all_comb$PriceHighSpeed)/all_comb$PriceHig
all_comb$diffmedtohigh <- (all_comb$PriceHighSpeed - all_comb$PriceMedSpeed)/all_comb$PriceMedSpeed * 10
```

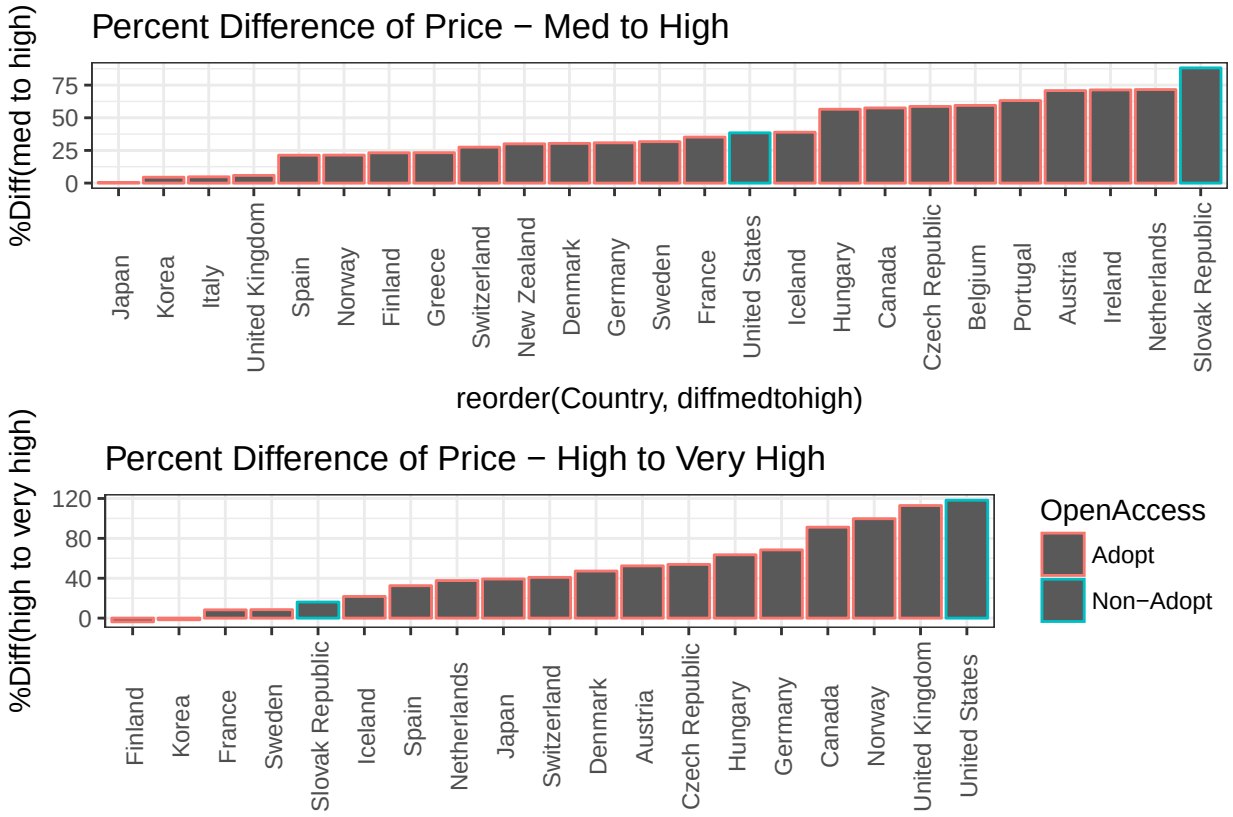
The newly defined columns are visualized to understand how it varies between countries that have adopted the open access policy vs. those which haven't.

```
# reorder the charts by value
diff1 <-all_comb %>%
  select(Country, OpenAccess, diffhightoveryhigh) %>%
  filter(!is.na(diffhightoveryhigh)) %>%
  ggplot(aes(x=reorder(Country, diffhightoveryhigh), y=diffhightoveryhigh, col=OpenAccess)) +
```

```
  geom_bar(stat="identity") + xlab("") + ylab("%Diff(high to very high)") + labs(title="Percent Differe
  theme(axis.text.x = element_text(angle=90, vjust=0.6), legend.position = "right") #+ coord_flip()

diff3 <-all_comb %>%
  select(Country, OpenAccess, diffmedtohigh) %>%
  filter(!is.na(diffmedtohigh) & ((diffmedtohigh > 0) & (diffmedtohigh < 100))) %>%
  ggplot(aes(x=reorder(Country, diffmedtohigh), y=diffmedtohigh, col=OpenAccess)) +
  geom_bar(stat="identity", show.legend = F) + ylab("%Diff(med to high)")  + labs(title="Percent Differe
  theme(axis.text.x = element_text(angle=90, vjust=0.6)) #+ coord_flip()

par(mfrow = c(1,2))  # 1 row and 2 columns
grid.arrange(diff3, diff1)
```



The plots above show the following about countries that have not adopted open access policies:

- Slovak Republic has the highest percent differenc in price from “Med” to “High”
- United States has the highest % difference in price from “High” to “Very High”
- Mexico doesn’t have data pertaining to “high” or “very high” category and so not in the plot (plot has filtered entries with NAs)

Countries such as Japan, Korea etc. show very small variations between the different tiers. Although this can be inferred from the data, further analysis is required using information on what is offered in these countries to understand what the benefits are from adopting to the open policies.

Conclusion

With the information provided on price, penetration and speed of broadband and wireless services across 30 countries, we can see how countries like Korea and Japan have really taken a big leap in comparison to

United States by making the services more available and affordable to its inhabitants. Apart from some outliers (or possibly bad data) identified with Luxemburg and Poland, we identified that 3G growth rate for a country like US in the dataset (value of 0) doesn't correlate with the 3G / household penetration information. While countries like Finland, Korea, Japan, Italy, France and Sweden show lower in price different tiers compared to United States, only Finland, Sweded and Italy show penetration rates that are substantially higher than United States. Japan, Korea and France are comparable to that of United States in 2G/3G penetration. Interestingly, Korea has the highest household penetration. Analysis of the speeds information shows that Korea offers the highest average download speeds, while Japan and Sweden are the leaders in avereage upload speeds. The weighted ranking plot shows how each country ranks overall and gives a good overall understanding in the spectrum of broadband services.

Based on the fact that certain countries have adopted open access policies while a few have not, it can be seen that there is a substantial drift in price between tiers among the countries that have not adopted the policy.