Trabajo final

David Cardona Duque

8/30/2021

Introduccion

El impacto del internet en las sociedades actuales es altamente visible a partir de los efectos en su construcción, por ejemplo, posibilitando la interconexión entre las personas o por otro lado generando obstáculos y brechas entre ellas ocasionando así implicaciones directas en el ejercicio de derechos y en la posibilidad de competitividad y progreso para las naciones, siendo entonces el Internet un objeto pertinente de estudio desde diversas perspectivas. Es desde allí que un acercamiento exploratorio toma relevancia para comprender y detectar distintas variables que atraviesan el desarrollo, capacidad, acceso y eficiencia respecto al internet, en especial, considerando su importancia como "fuerza impulsora de la aceleración de los progresos hacia el desarrollo en sus distintas formas" (Asamblea General de las Naciones Unidas, 2012, p.2).

Objetivo

Construir análisis e hipótesis respecto a la variación de velocidad del internet con base en la lectura de los resultados obtenidos por medio de la exploración multivariada, bivariada y univariada.

Metodologia

- Análisis exploratorio desde los siguientes procedimientos:

- Selección de variables pertinentes para el análisis según las suministradas por el curso y la observación propia del fenómeno.

- Recolección de datos por medio de páginas web y bases de datos pertinentes para las variables escogidas.

- Depuración de los datos mediante el software R y Rstudio según la identificación de variables lejanas a la lectura promedio de la realidad y diferencias entre las bases de datos.

- Análisis exploratorio por medio de preguntas claves relacionando las distintas variables seleccionadas.

- Acercamiento a hipótesis a partir de la lectura de los resultados obtenidos del análisis multivariado, bivariado y univariado.

Depuración.

Sobre la base de datos "internet1"

Se cambia el tipo de la variable "TEST_DATE" a fecha puesto que es el tipo real de esta, se ignora los minutos debido a que para el objetivo del análisis no son relevantes porque se tiene en cuenta la hora de toma en general para comparar las velocidades.

Se cambia el nombre de la variable "SERVER_NAME" por "SERVER_LOCATION", ya que se acerca más al significado real de la misma, pues sus datos son las ciudades de locación de los servidores.

Se cambia el nombre de la variable "nombre" por "SERVER_NAME", ya que se acerca más al significado real de la misma, pues sus datos son los nombres de las empresas operadoras de los servidores.

Se unen las bases de datos "internet1" e "internet2" para realizar análisis a posteriori y se guardan en una nueva base llamada "internet"

Sobre la base de datos "internet2".

Se cambia el tipo de la variable "TEST_DATE" a fecha puesto que es el tipo real de esta, se ignora los minutos debido a que para el objetivo del análisis no son relevantes porque se tiene en cuenta la hora de toma en general para comparar las velocidades.

Se cambia el nombre de la variable "SERVER_NAME" por "SERVER_LOCATION", ya que se acerca más al significado real de la misma, pues sus datos son las ciudades de locación de los servidores.

Se cambia el nombre de la variable "nombre" por "SERVER_NAME", ya que se acerca más al significado real de la misma, pues sus datos son los nombres de las empresas operadoras de los servidores.

Sobre la base de datos "internet3".

Se cambia el tipo de la variable "TEST_DATE" a fecha puesto que es el tipo real de esta, se ignora los minutos debido a que para el objetivo del análisis no son relevantes porque se tiene en cuenta la hora de toma en general para comparar las velocidades.

Se cambia el nombre de la variable "SERVER_NAME" por "SERVER_LOCATION", ya que se acerca más al significado real de la misma, pues sus datos son las ciudades de locación de los servidores.

Se cambia el nombre de la variable "nombre" por "SERVER_NAME", ya que se acerca más al significado real de la misma, pues sus datos son los nombres de las empresas operadoras de los servidores.

Sobre la base de datos "internet".

Se unen las bases de datos "internet" e "internet3" para realizar análisis a posteriori y se guardan en una nueva base llamada "internet_speed".

Sobre la base de datos “internet_speed”.

Se hace una conversión de unidades a kilómetros en la variable “DISTANCE_MILES” la cual estaba en millas, porque esta unidad de medida es mucho más común y aceptada mundialmente.

Se cambia el nombre de la variable “DISTANCE_MILES” por “DISTANCE_KM”, ya que se acerca más al significado real de la misma, pues sus datos simbolizan la distancia en kilómetros del lugar de toma de datos al lugar de ubicación del servidor.

Se crea una nueva variable llamada “perdida_download_porcentaje” la cual represente el porcentaje de perdida de la velocidad de internet de descarga medida vs. la contratada.

Se crea una nueva variable llamada “perdida_upload_porcentaje” la cual represente el porcentaje de perdida de la velocidad de internet de subida medida vs. la contratada.

Todos los valores negativos de la variable “perdida_download_porcentaje” se cambian por 0, pues estos no tienen valor descriptivo cuando son negativos, ya que nos interesa saber el porcentaje de perdida de la velocidad de descarga.

Todos los valores negativos de la variable “perdida_upload_porcentaje” se cambian por 0, pues estos no tienen valor descriptivo cuando son negativos, ya que nos interesa saber el porcentaje de perdida de la velocidad de subida.

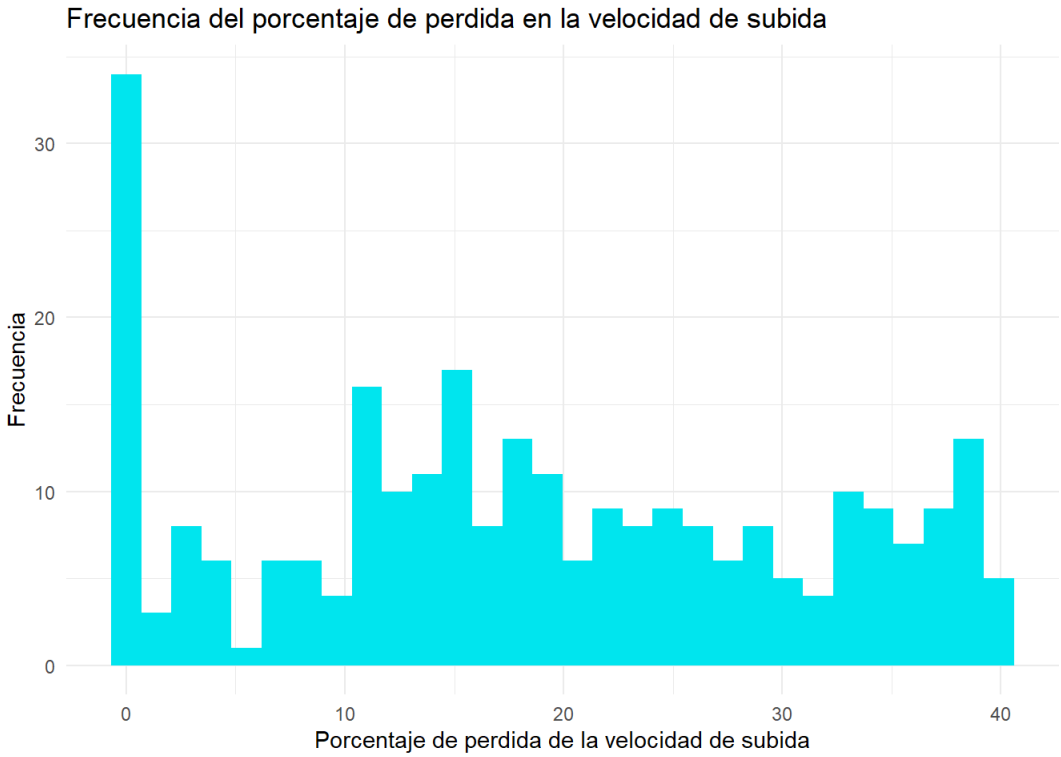
Análisis descriptivo y exploratorio

Univariado

Velocidad de subida

Summary porcentaje perdida subida

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.03   17.65   18.57   28.85   39.90
```

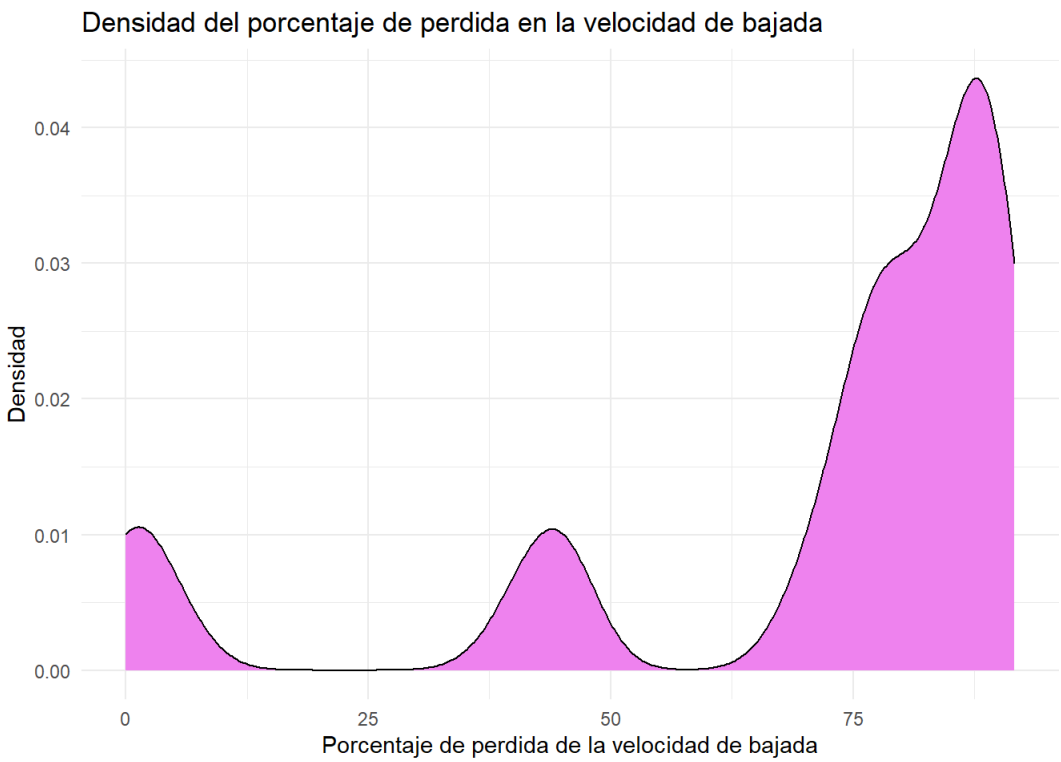


Desde el gráfico y la tabla resumen podemos deducir que en el caso de la velocidad de subida la perdida es mínima en comparación con la de bajada, se puede observar claramente que a comparación de los otros valores existe uno muy predominante que tiene un valor extremadamente bajo. Cuando analizamos la tabla resumen descubrimos que la el promedio de perdida es de 18.57 porciento. Todo esto debido posiblemente a que la capacidad de subida de una red en general es menos usada que la capacidad de bajada gracias a que en general esta velocidad se utiliza para actividades más cotidianas como descargar archivos, reproducir vídeos en Streaming, entre otros. Por lo tanto se tiene más capacidad disponible de manera individual en el dispositivo a la hora de hacer la prueba.

Velocidad de bajada

Summary porcentaje perdida bajada

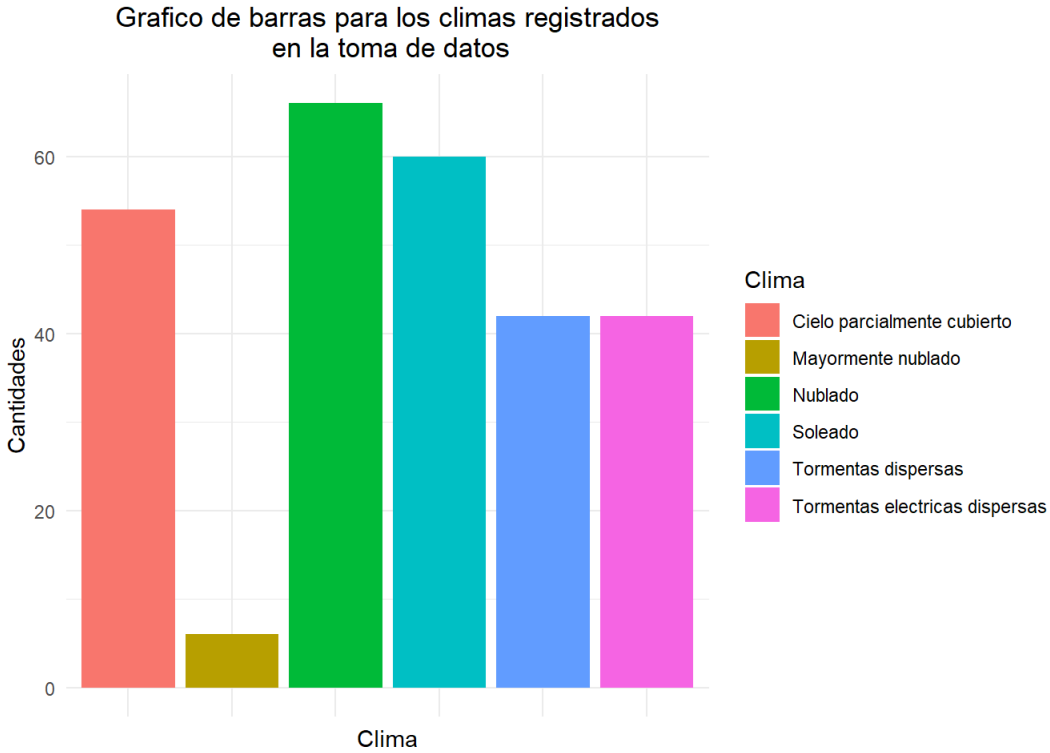
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   71.39   79.73   69.79   88.00   91.63
```



Desde ambas representaciones de los datos se puede percibir una alta concentración de los datos en los de perdida altos, desde la tabla el promedio es de 69.79 el cual es considerable. Si observamos en comparación a la velocidad de subida la perdida es preocupante, una hipótesis para esto es que esta velocidad es altamente utilizada por la generalidad de las personas para su vida diaria, esto sumado a las distancias de los servidores promueve una perdida importante. Por otro lado, existen 2 acumulaciones además de la principal, una en valores bajos y otra en medios, una posible razón para esto es que dado a la distancia de los servidores existen casos donde la perdida en general no es tan alta porque los datos llegan al dispositivo que realiza la prueba de una manera más integra.

Clima

```
##
##   Cielo parcialmente cubierto            Mayormente nublado
##                            54                             6
##                       Nublado                       Soleado
##                            66                            60
##          Tormentas dispersas Tormentas electricas dispersas
##                            42                            42
```



El dato más común dentro de la base de datos en la variable clima es claramente nublado, pero esto no termina siendo preocupante para análisis futuros en términos de información sesgada, ya que no es un dato predominante en la generalidad y el dato que le sigue en frecuencia, el cual es soleado, es transversalmente distinto en cuanto a condiciones, por lo tanto si se cuenta con un buen campo de análisis a la hora de generar hipótesis.

Bivariado

Relacion Clima y porcentaje de perdida en la velocidad de bajada

summary clima Cielo parcialmente cubierto

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   73.75   79.58   71.96   87.53   91.37
```

summary clima Mayormente nublado

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   44.07   53.77   77.98   67.64   78.28   81.73
```

summary clima Nublado

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   72.97   81.78   69.91   88.20   91.10
```

summary clima Soleado

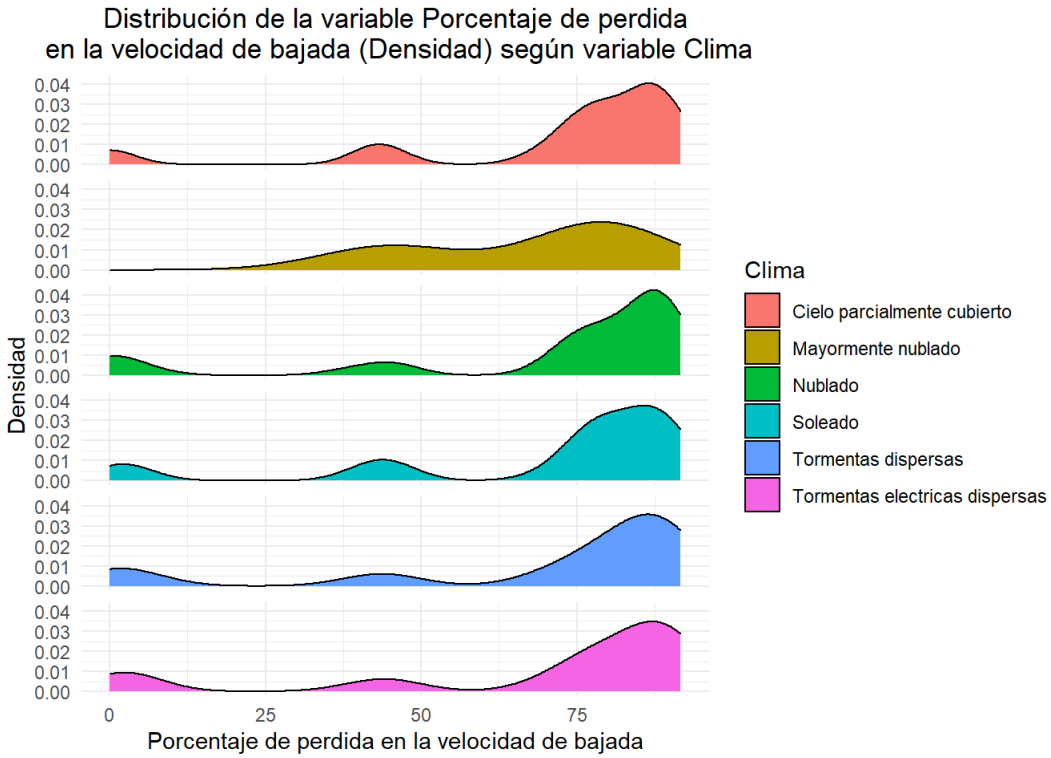
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   71.52   79.05   69.70   87.37   90.80
```

summary clima Tormentas dispersas

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   69.52   81.75   68.40   88.38   91.43
```

summary clima Tormentas electricas dispersas

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   70.72   80.62   68.62   88.98   91.63
```



Desde las dos representaciones de los datos se puede deducir que el cambio entre los distintos tipos de clima es realmente mínimo, esto se evidencia porque la diferencia entre el promedio más bajo que es el del clima mayormente nublado y el más alto que es el de clima parcialmente cubierto es de 4.32 porciento. Una hipótesis para esto es que realmente las formas de transmisión de datos alrededor del mundo no dependen del clima pues estos se transmiten con cables submarinos a través de los continentes y luego mediante cables de los proveedores hasta los el sitio de toma de la prueba a los cuales poco o nada les afectaran estas condiciones.

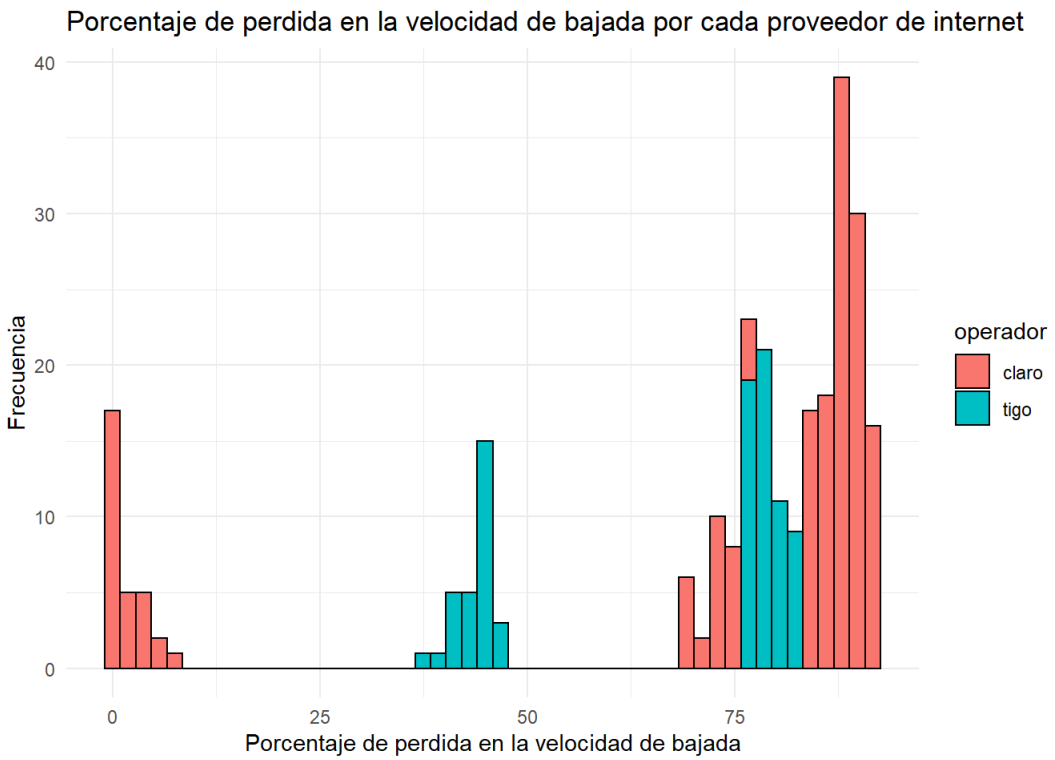
Relacion Operador y porcentaje de perdida en la velocidad de bajada

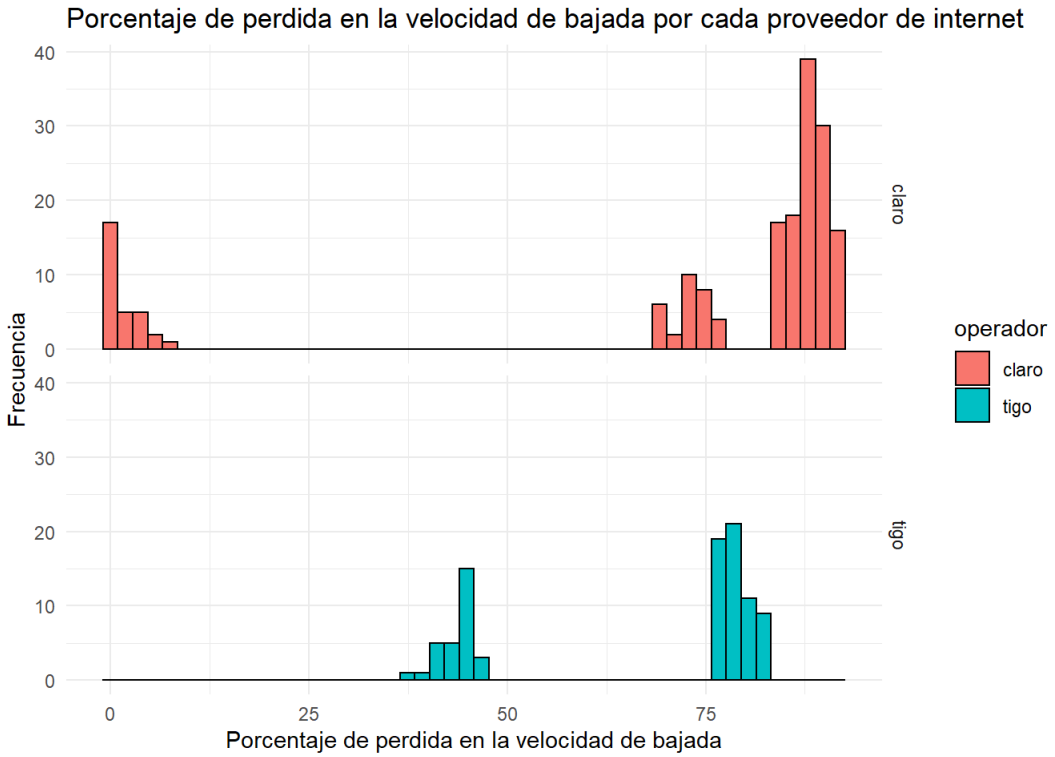
summary perdida en la velocidad de descarga operador claro

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   73.62   86.53   71.12   88.88   91.63
```

summary perdida en la velocidad de descarga operador tigo

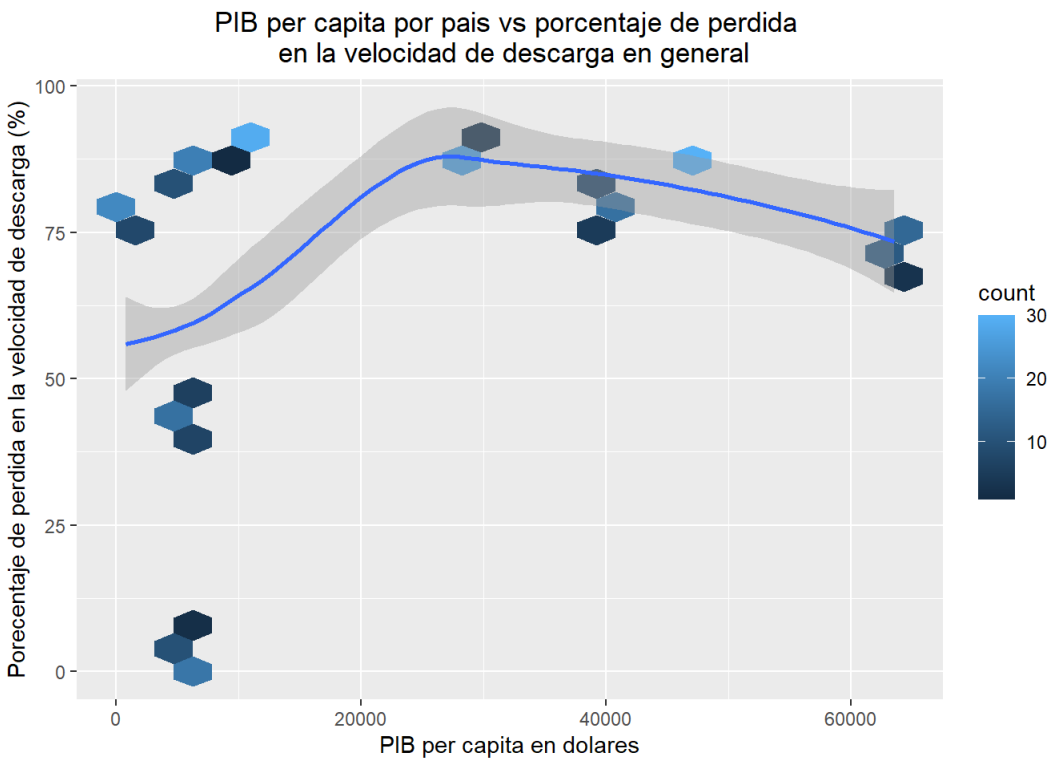
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   38.07   45.35   77.47   67.13   79.22   83.20
```





Desde un análisis exploratorio de ambas graficas y de las tablas de resumen se puede afirmar que existe una diferencia mínima entre el porcentaje de perdida de la velocidad de descarga y el operador de internet, siendo el operador tigo el cual tiene una mayor fidelidad de su velocidad. La poco diferencia entre ambos puede ser debido a que en general ambas compañías de internet poseen una infraestructura eficaz a la hora de prestar el servicio de internet aunque en general por la ciudad en la cual se tomaron los datos, la cual fue Medellín, se tiene un mayor afianzamiento del operador tigo porque este es producto de una unión con UNE el cual era el proveedor principal de esta ciudad anteriormente gracias a que pertenecía a las Empresas Públicas de Medellín (EPM).

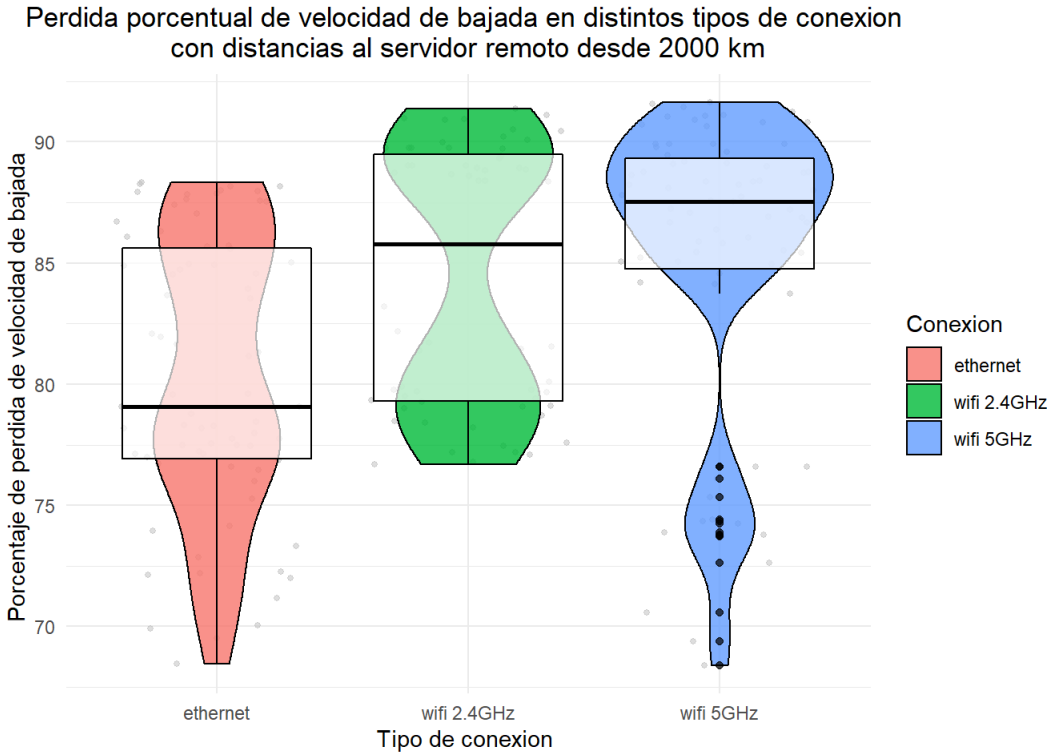
Relacion entre el PIB per capita del pais donde esta el servidor remoto y el porcentaje de perdida en la velocidad de bajada



Antes de esta gráfica se partía de un supuesto lógico el cual era que a mayor PIB per cápita se tenía una menor perdida en la velocidad de internet, puesto que se tenían recursos en los países donde están los servidores para proveer una mejor calidad de red, luego de analizar esta misma nos dimos cuenta de que en cierto el PIB per cápita termina aumentando esta perdida, esto posiblemente debido a que en estos lugares se genera un mayor uso local de la red de internet, pues las personas que viven allí tienen un mayor poder adquisitivo y por la tanto más accesos a este bien, luego observamos un punto de inflexión donde a mayor PIB per cápita la perdida comienza a disminuir, podríamos explicar esto debido a que en este punto el poder adquisitivo de las personas se ve superado por el avance en la red general pues el país tiene los recursos suficientes y comienza a generar un beneficio para pruebas remotas. Por otro lado, vemos como hay observaciones específicas en un PIB per cápita especialmente alto, esto debido a que estas lecturas son las que se generaron contra un servidor extremadamente cerca al usuario y por lo tanto se garantiza la velocidad de bajada.

Multivariado

Cuándo se tiene una distancia alta (desde 2000 KM) al servidor remoto, ¿cómo es el comportamiento del porcentaje de perdida de la velocidad de descarga en los diferentes modos de conexión?



summary del porcentaje de velocidad de descarga perdido con la conexión vía Ethernet cuando hay una distancia desde 2000 KM

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   68.45   76.93   79.07   80.37   85.62   88.32
```

summary del porcentaje de velocidad de descarga perdido con la conexión vía wifi 2.4 GHz cuando hay una distancia desde 2000 KM

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   76.70   79.32   85.78   84.44   89.49   91.37
```

summary del porcentaje de velocidad de descarga perdido con la conexión vía wifi 5 GHz cuando hay una distancia desde 2000 KM

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   68.38   84.77   87.53   85.21   89.32   91.63
```

Este gráfico es una combinación de un diagrama de cajas y bigotes y un diagrama de densidad girado y colocado a cada lado, para mostrar la forma de distribución de los datos. Así del mismo podemos obtener que la forma de conexión cuando se tienen distancias altas que menos perdida de velocidad de bajada tiene es el Ethernet, o sea, conexión por cable, teniendo este su grueso de datos en 2 partes especificas: los rangos desde 75% hasta 80% y los rangos de 85% hasta 90%, cosa la cual no es sorpresa alguna porque se parte que existe una distancia alta entre el servidor y el equipo que está haciendo la prueba remota y además que la conexión por cable en comparación a las otras 3 tiene generalmente una mejor calidad de red. Ahora analizando las 2 restantes si nos llevamos una sorpresa, el Wifi de 2.4 GHz tiene un mejor desempeño que el WiFi de 5.0 GHz a pesar de que suele tener más interferencias haciendo que su conexión vaya un poco más lento de lo que debería. Por otro lado, el WiFi 2.4 GHz tiene menos canales debido a las restricciones que existen respecto a estos ocasionando que a cada dispositivo le toque menos capacidad de red. Una posible razón para este fenómeno es que la tecnología WiFi 5 GHz no tiene un avance necesario en la ubicación donde está el equipo con el que se toma la prueba para mostrar sus beneficios, además que gracias a la narrativa de que este es mejor, más personas se han estado pasando a este y por lo tanto se genera una mayor interferencia en la frecuencia de 5 GHz, además que se descongestiona la frecuencia de 2.4 GHz. Aunque es importante notar que a pesar de lo anterior, en la gráfica correspondiente a 5 GHz hay una cantidad importante de datos atípicos, señalado por la frecuencia azul mostrada en la parte baja de la misma, esto es claramente una consecuencia de que el WiFi 5 GHz si tiene un potencial importante al implementarse de manera correcta. En conclusión el Ethernet es la mejor forma de conexión para minimizar el porcentaje de perdida en la velocidad de descarga, seguido a esto observamos como actualmente la frecuencia de 2.4 GHz termina siendo mejor a pesar de ser una tecnología más antigua que la de 5 GHz, aunque se observa una mejora significativa, casi superando al, Ethernet, en casos atípicos de la misma, lo que demuestra entonces su potencial con un mejor desarrollo.

Cuándo se tiene una cantidad de dispositivos conectados alta (desde 6), ¿cómo es el comportamiento del porcentaje de perdida de la velocidad de descarga en las diferentes etapas del día?

summary de el porcentaje de velocidad de descarga perdido en la etapa del dia mañana cuando hay desde 6 dispositivos conectados

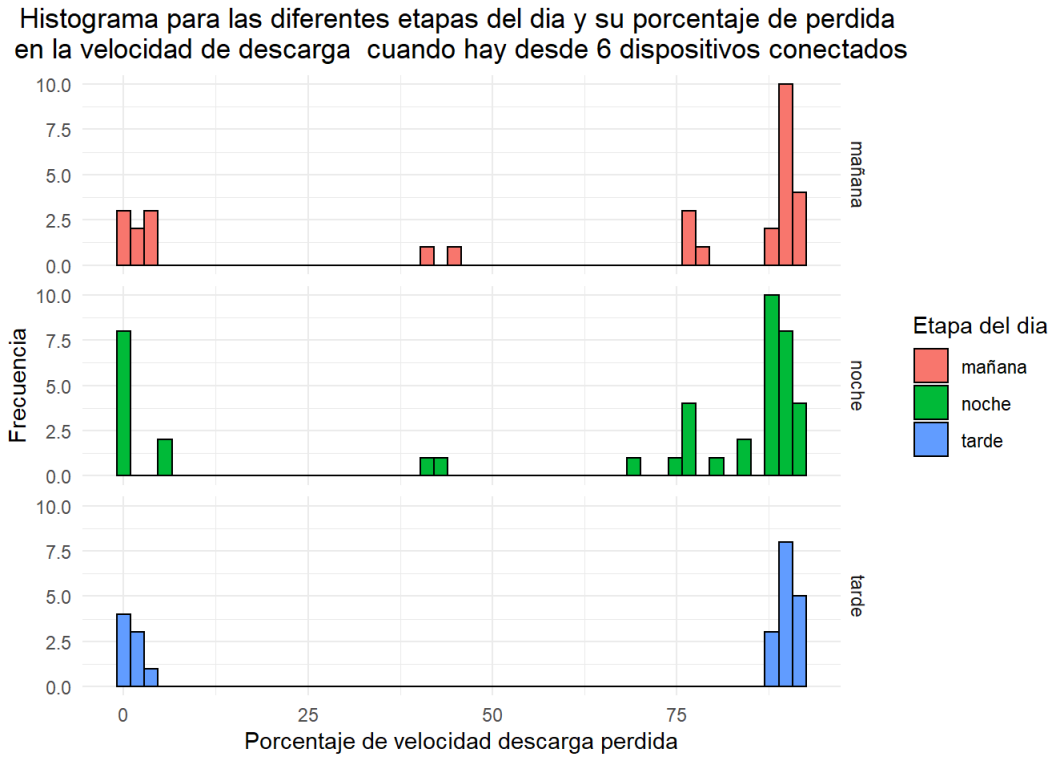
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   13.43   88.50   61.62   89.79   91.63
```

summary de el porcentaje de velocidad de descarga perdido en la etapa del dia tarde cuando hay desde 6 dispositivos conectados

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.425  88.933  60.368  90.046  91.233
```

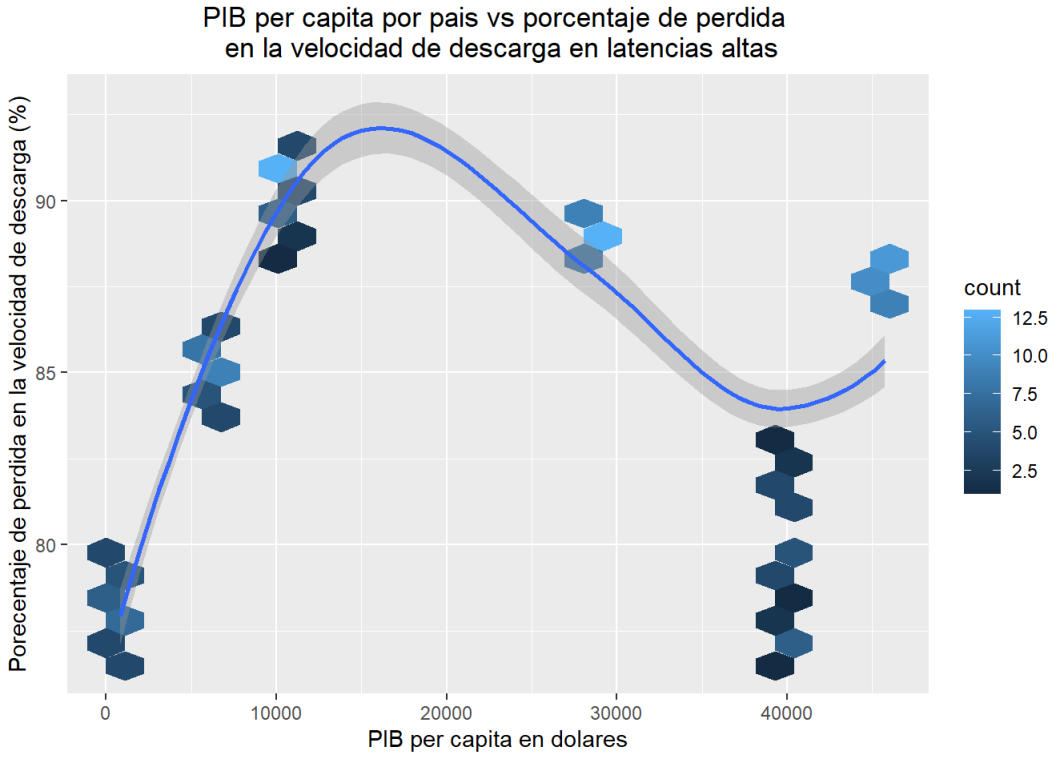
summary de el porcentaje de velocidad de descarga perdido en la etapa del dia noche cuando hay desde 6 dispositivos conectados

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   41.88   87.05   64.27   89.11   91.43
```



Desde los histogramas se puede interpretar que cuando hay una gran cantidad de dispositivos conectados (más de 6) el porcentaje de perdida de la velocidad de descarga del internet es en todas las etapas del día o muy alto o muy bajo, existiendo extremadamente pocos datos intermedios, esto debido posiblemente a que cuando una red se somete a una gran carga de trabajo las cosas se ven extrapoladas a 2 casos específicos y son: un muy buen desempeño de la misma, lo cual podemos ver en la gráfica que es en pocas ocasiones, dado normalmente por una buena infraestructura de red como por ejemplo fibra óptica, o en su defecto una perdida de velocidad de descarga alta, lo cual podemos ver como lo más común, que sucede cuando se tiene una infraestructura de red normal. En otro orden de ideas, el análisis comparativo de las 3 etapas del día nos lleva a descubrir que la etapa con menos perdida de internet es la tarde, lo cual se confirma con la tabla resumen pues es la que tiene un promedio menor. Una hipótesis para lo anterior es que en las tardes se desarrollan comúnmente las actividades con una mayor necesidad de internet, como por ejemplo el trabajo desde casa y el estudio, por lo tanto es probable que desde las compañías de internet se tenga un esfuerzo mayor para una calidad del servicio, puesto que deben garantizar un internet decente al grueso de sus clientes, en comparación con la noche y la mañana cuando las actividades de uso de internet fuertes no son tan frecuentes y en este caso aún se tiene una gran cantidad de dispositivos conectados lo que conlleva a una mayor perdida de la velocidad de descarga debido a una falta de incentivo a una atención especial por parte de las compañías de internet. En conclusión cuando se tiene una cantidad de dispositivos alta existen 2 casos principales: que se tenga una perdida alta en la velocidad de descarga o que esta perdida sea extremadamente baja, además que la etapa del día con menor perdida es la tarde debido asumiblemente a que se tiene un mayor esfuerzo de parte de los proveedores de internet para tener un servicio de calidad.

¿Existe alguna relación entre el PIB per cápita del país en el que está el servidor de prueba y el porcentaje de perdida de la velocidad de descarga cuando se tienen latencias altas (desde 150 ms)?



Cuando se tiene una latencia alta, o sea un tiempo de respuesta alto del servidor, se puede observar en la primera gráfica que en los países donde se ubican los servidores cuando se tiene un mayor PIB existe una relación ambigua con el porcentaje de perdida de velocidad de descarga en comparación con la contratada, se ve como en principio a mayor PIB la perdida de velocidad aumenta significativamente, una hipótesis para esto es que cuando se parte de una latencia alta esto implica que existe un problema en sí mismo con la conexión al servidor y un mayor PIB per cápita en esta parte lo agrava porque existe un mayor poder de adquisición por persona, por lo tanto, es probable una mayor posesión de aparatos electrónicos que utilicen el internet lo que genera una mayor carga de dispositivos a estos servidores, haciendo así menos efectiva la conexión por parte de externos remotamente. Por otro lado, existe un punto de inflexión donde el aumento del PIB comienza a mejorar la perdida de velocidad, esto posiblemente porque en este punto el ingreso per cápita del país donde se encuentra el servidor, al ser considerablemente mayor, permite un mejoramiento general de la red de internet y por lo tanto una mejor velocidad en conexiones externas al tener más poder de procesamiento. Otro punto de importancia es el caso específico de Tokio - Japón, al cual pertenecen los datos más bajos luego del punto de inflexión, esta bajada drástica es debido a que en este caso especial, Japón ha invertido una cantidad considerable de dinero en el mejoramiento de su red de internet, la cual es mucho mayor en comparación a la de otros países con su mismo o más PIB per cápita gracias a que se desarrollaron los últimos juegos olímpicos en Tokio y por lo tanto se necesitaba una calidad de internet desarrollada para un correcto desarrollo, por ejemplo, en la mayoría de las competencias se tenía una transmisión a las familias de los deportistas. En conclusión en este análisis se descubrió que en latencias altas un mayor PIB per cápita implica una mayor perdida en la velocidad de descarga del internet hasta un punto de inflexión donde el PIB permite generar una mejora en las infraestructuras de conexión y por lo tanto una menor perdida de velocidad de descarga en el enlace remoto.

Conclusiones

- La velocidad de subida no tiene casi perdida en cuanto a la velocidad contratada, posiblemente porque esta velocidad no es usada tan masivamente por las personas en la vida diaria y por lo tanto a la hora de usarla hay mayor capacidad de red disponible.

- El tipo de clima poco afecta a la velocidad de descarga, esto hipotéticamente porque la forma en la que se transmite el internet no se ve afectada por estas condiciones climáticas

- El operador de internet tiene una baja relevancia a la hora de evaluar la velocidad de descarga, aunque en el caso específico de Medellín se tiene que tigo posee una ligeramente mejor red.

- El PIB per cápita afecta de una forma ambigua al porcentaje de perdida de velocidad de descarga en general, teniendo un punto de inflexión donde la comienza a mejorar.

- El Ethernet es la mejor forma de conexión para minimizar el porcentaje de perdida en la velocidad de descarga, seguido a esto observamos como actualmente la frecuencia de 2.4 GHz termina siendo mejor a pesar de ser una tecnología más antigua que la de 5 GHz, aunque se observa una mejora significativa, casi superando al, Ethernet, en casos atípicos de la misma, lo que demuestra entonces su potencial con un mejor desarrollo.

- Cuando se tiene una cantidad de dispositivos alta existen 2 casos principales: que se tenga una perdida alta en la velocidad de descarga o que esta perdida sea extremadamente baja, además que la etapa del día con menor perdida es la tarde debido asumiblemente a que se tiene un mayor esfuerzo de parte de los proveedores de internet para tener un servicio de calidad.

- En latencias altas un mayor PIB per cápita implica una mayor perdida en la velocidad de descarga del internet hasta un punto de inflexión donde el PIB permite generar una mejora en las infraestructuras de conexión y por lo tanto una menor perdida de velocidad de descarga en el enlace remoto.

Recomendaciones

- Tener en cuenta en los análisis respecto al Internet los distintos tipos de estructura de red que coexisten en la actualidad, en especial, debido a la etapa de cambio que se está desarrollando, permitiendo así tener hipótesis más cercanas a la realidad.

- Diversificar en mayor medida la cantidad de observaciones sobre la variable clima para tener un mejor rango experimental, su importancia debido a que es una variable de la cual no se tiene control de escogencia en la toma de datos.

- Emplear más variables de respuesta para tener un análisis más holístico, debido a que este es fundamental en un estudio exploratorio con el fin de evitar la omisión de características de interés.

Blibliografia

Asamblea General de las Naciones Unidas. (2012, 29 junio). Promoción y protección de todos los derechos humanos, civiles, políticos, económicos, sociales y culturales, incluido el derecho al desarrollo [20° periodo de sesiones]. https://ap.ohchr.org/documents/S/HRC/d_res_dec/A_HRC_20_L13.pdf

Grupo Banco Mundial. (s. f.). PIB per cápita (US$ a precios actuales) | Data. Banco Mundial. https://datos.bancomundial.org/indicador/NY.GDP.PCAP.CD

Provincia China de Taiwán PIB per cápita, 1980–2020. (s. f.). Knoema. https://knoema.es/atlas/Provincia-China-de-Taiw%C3%A1n/PIB-per-c%C3%A1pita

Oficina de Estrategia Internacional de Información y Comunicación. (2016, enero). Política japonesa de información y comunicaciones en la era de Big Data (datos masivos) e IoT (Internet de las cosas) - Hacia el año 2020. MIC. https://www.omc.co.jp/ICTperu/espanol/pdf/01.pdf

Ookla. (s. f.). Speedtest by Ookla - The Global Broadband Speed Test. Speedtest.Net.https://www.speedtest.net/