Data Import and Cleaning

1. Loaded all required libraries.
2. Using fread () function from package data.table to read the dataframe’s, for using optimal memory storage.
3. All 3 csv’s are read into person, download and upload dataframe respectively.
4. We are now creating a table that has
   * The structure ***`person\_id, city, type\_of\_broadband\_connection, name\_of\_isp, average\_download\_speed, average\_upload\_speed`.***
   * ***has 1 line per person (i.e. calculate a single average download and upload speed for each person)***
   * ***only contains people in the cities 'Samsville' and 'Databury'***
   * ***only contains download and upload measurements that have run successfully (i.e. put a filter on did\_test\_complete\_successfully)***
   * ***only contains tests run in the month of January 2021 (i.e. put a filter on time\_of\_measurement).***
5. For this we have considered first ‘download’ and ‘upload’ dataframe in which we have filtered the column ***did\_test\_complete\_successfully == TRUE*** and ***time\_of\_measurement = January 2021*** *a*nd stored into new dataframe ‘**download\_filtered**’ and ‘**upload\_filtered’**.
6. Now taking person dataframe, we have filtered based on city ***'Samsville' and ‘Databury’ and*** joined all 3 dataframe based on ‘person\_id’. With summarizing the values and calculating ‘**average\_download\_speed’** and ‘**average\_upload\_speed’.**
7. We have the dataframe ‘***df***’ ready with the above structure.

Calculate the average download speed each day, and take the 60th percentile

1. To calculate the average download speed each day we look into ‘**download**’ dataframe.
2. We filtered the download dataframe applying group\_by function with respect to ‘**person\_id**’ and ‘**time\_of\_measurement**’. And summarizing average of column ‘**measured\_download\_speed\_in\_Mbps**’ and storing in new column named ‘**average\_download\_speed**’ and stored into new dataframe ‘**df\_daily\_speeds**’.
3. With dataframe ‘**df\_daily\_speeds**’ we calculated 60 percentile of average download speed column and stored in new dataframe ‘**df\_percentile’** which defines 60th percentile of each user which will help to find out the answer for question **'on at least 60% of days, average download speed for user (person\_id) was x (60th percentile value)**’.
4. Finally we joined the df\_percentile dataframe with df and created ‘**final\_table**’ dataframe.

Data Quality

In Data Quality we are checking some basic descriptive statistics of dataframe. We need to make sure the data we are using is correct and proper to provide meaningful insights and reports.

Like any dataset, this one has problems; for example:

\* There are people in different cities

\* People who have differing rates of missing data

\* There are people with average speeds that are either too high or too low for their `type\_of\_broadband\_connection` (i.e. mislabelled data)

\* There are people who have OK data overall, but some individual tests fall outside of parameters.

We are highlighting the above points to see errors/ missing values/ mislabelled data in our dataframe

1. We are performing some basic descriptive statistics using functions **dim ()** that indicates dimension of dataframe, **diagnose ()** shows data types, missing values, missing values percentage, unique count etc., **describe ()** shows basic business moment decisions such as mean, variance, standard deviation, IQR, min, max for all the dataframe.
2. We plotted a BarChart that shows the count of people in each city.
3. We have checked the outliers for ‘measured\_download\_speed\_in\_Mbps’, ‘measured\_upload\_speed\_in\_Mbps’, ‘average\_download\_speed’, ‘average\_upload\_speed’ using Boxplot.
4. We have encountered some values in download and upload dataframe where for the column ‘did\_test\_complete\_successfully’ value is ‘TRUE’ but the respective speed calculated is 0. This can be considered as Error test value.
5. We have created a dataframe ‘sorted\_df’ which depicts the count of error test values for each person. We have some meaningful insights from this dataframe which are
   1. person\_id 96610 has highest number of error test value with 174 and person\_id 38202, 49957 and 89731 has the lowest error test value with 1.
   2. The ‘Fibre’ broadband connection provides very low error test values i.e., less than equal to 12 and ‘VDSL’ provides very high error test values with more than 130+ for almost 50 users.
   3. All the error test value reported for ‘VDSL’ broadband connection is provided by ‘Useus’ ISP provider.
6. The 2 barplot shows the chart of count of error test with value 0 by broadband connection and sum of error test with value 0.
7. We have also identified the mislabelled data, as we know the download speed provided by each broadband ‘ADSL’, ‘VDSL’, and “Fibre’ is 10, 100 and 1000 Mbps respectively. So we have categorized the lower and higher value for each broadband and assigned to each broadband value.
8. With ifelse () condition we are checking the parameters of mislabelled broadband connection with average download speed and id the value is within range it prints ‘Yes’ and if not it print ‘No’
9. And finally with barplot we have seen that there is no mislabelled data for ‘VSDL’. However there are some mislabelled data in ‘Fibre’ and most mislabelled data in ‘ASDL’.

Data summarization and presentation

Let’s summarize the dataframe for download and upload speed to see which ISP is providing good download speed with which particular Broadband and in which city.

1. download\_summary dataframe provides the average speed provided by ISP’s with different broadband connection and over different cities. As we can see Fibrelicious tends to provide higher average download speeds compared to Useus across all connection types (ADSL, VDSL, and Fibre) in both Samsville and Databury. Except for ‘ADSL’ in Samsville which is lower than Useus.
2. The difference in average download speeds between the two ISPs is most noticeable in the Fibre connection type. This suggests that Fibrelicious may have a competitive advantage in terms of faster download speeds, particularly for Fibre connections.
3. Whereas in upload summary dataframe the Useus tends to provide good upload speed for ‘Fibre’ connection in both cities. The ‘ADSL’ provide good average upload speed with Fibrelicious in Databury but Useus provide better average upload speed in Samsville.
4. However we are making this statement on the average data that is collected, and this is not reflected by any other factors such as network, pricing, customer service etc.
5. Further we have plotted a bar chart to of average download speed provided by ISP’s with respect to city and type of broadband connection. And from chart we can see in ASDL provided by Fibrelicious in Databury city there is some dirty data as it is showing the average speed more than 10 Mbps.
6. We have also compared the average download speed provided by both the ISP’s and summarized that ‘**If you have a Fibre connection in Databury, Fibrelicious provides a better download speed by approximately 31.54, 33.55, 41.52, 27.64, 30.07, 201.01, 28.77, 34.11, 34.03, 170.75, 26.76, 35.13, 33.06, 30.54, 32.31, 33.51, 34.8, 26.75, 32.7, 31.77, 32.49, 27.17, 35.44, 40.01, 28.66, 29.28, 21.48, 27.64, 34.43, 34.79, 170.01, 25.34 Mbps compared to Useus’.**
7. And finally showing comparison of average download speed per hour by different ISP’s. we can see that Fibrelicious provide better average download speed up till 17:00 and Useus provide better speed than Fibrelicious between 18:00 to 22:00 period of time