CWE CONNECTING WINDSOR-ESSEX

ST. CLAIR COLLEGE

Data Assessment

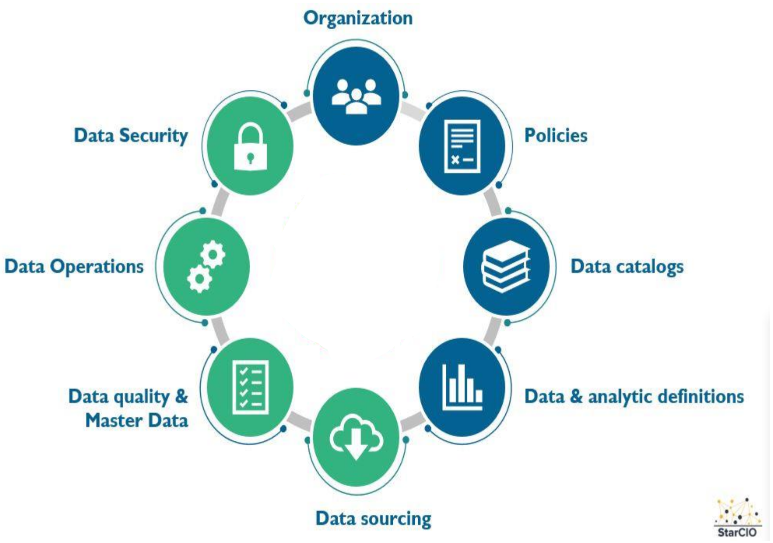
Capstone Project: Connecting Windsor-Essex: Broadband

The objective will be in regard to provide analysis and recommendations concerning broadband using the data provided by CW-E stakeholders and the services currently deployed by various school boards in Southwestern Ontario. These insights include, but are not limited to:

- Analysis of broadband quality
- Analysis of broadband speed
- Comparisons between carriers at the same location
- Comparisons between different technologies (i.e. fibre, coax, cellular, satellite)
- Comparisons between different sites with the same carrier/technology
- Analysis of external factors (i.e. weather) on internet services.

Data Assessment Steps:

Assessing the provided data by stakeholders against rules specified by them. Assess data against multiple dimensions such as accuracy of key attributes, completeness of all required attributes, consistency of attributes across multiple data sets, timeliness of data, etc. Depending on the volume and variety of data and the scope of Data Quality project in each phase, we might perform qualitative and/or quantitative assessment using some profiling tools. This is the stage to **assess policies as directed by the stake holders related to school boards and CW-E** (data access, data security, adherence to specific school board standards/guidelines, etc.) as well.



1. **Data Governance**: Data governance refers to the overall management of data as a strategic asset within an organization or project.

   - In the context of our **CW-E: Braodband project**, data governance factors including **Consent, Clarity, Consistency, control, Consequences** ensures that our data is properly collected, stored, processed, and used in a manner that supports the project goals and objectives as mentioned by the stakeholders.

   - This includes establishing policies, procedures, and standards for data management, as well as **ensuring compliance with relevant laws and regulations set by CW-E and St.Clair College stakeholders**.

   - The goal of data governance in a project is to ensure that data is of high quality, secure, and usable, while also promoting collaboration and consistency across all project stakeholders. **Effective data governance helps to minimize risks, improve data-driven decision making, and increase the overall success of the project.**

2. **Definition:** Defining the business goals for data quality improvement, data owners/stakeholders relating to CW-E, St. Clair College and respective school boards and their data rules.

- Framing questions with respect to our CW-E Broadband project is most important part of identifying business objective, for example:
   1. What is Connecting Windsor-essex organisation's overall objectives?

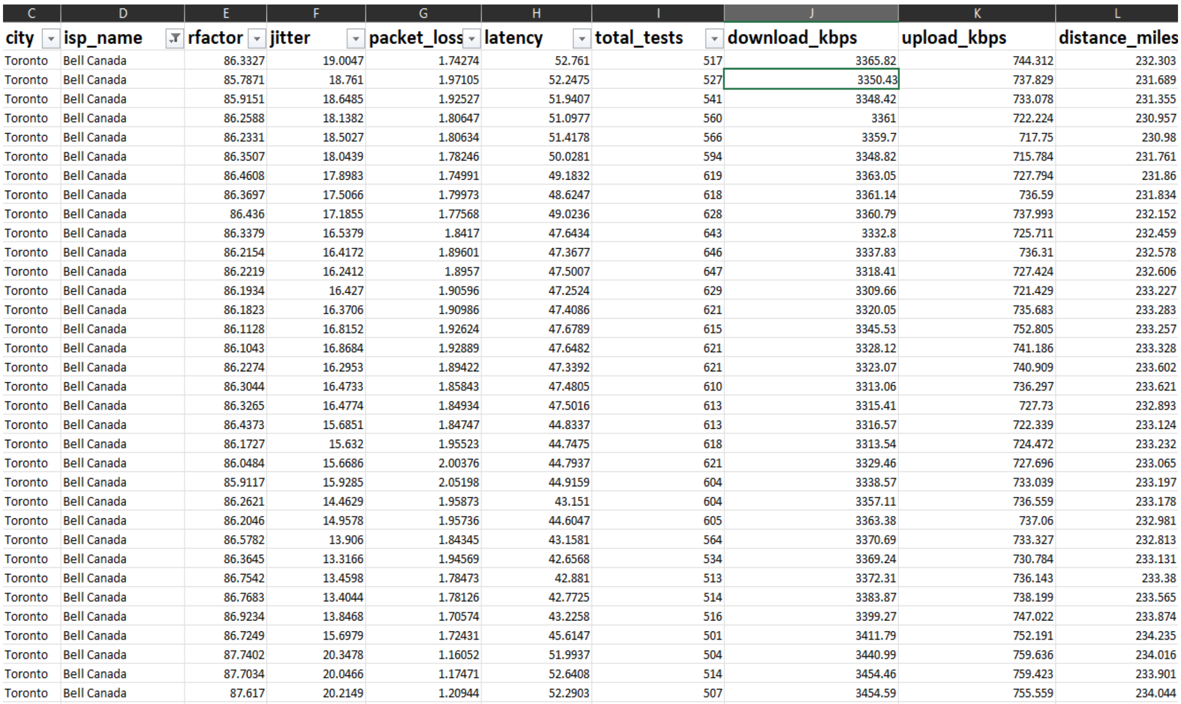
   2. What data is needed to meet these objectives?

   3. What types of insights and information are required to make progress against these initiatives?

   4. How does variables like Jitter, Latency, weather, Distance affect broadband speed.

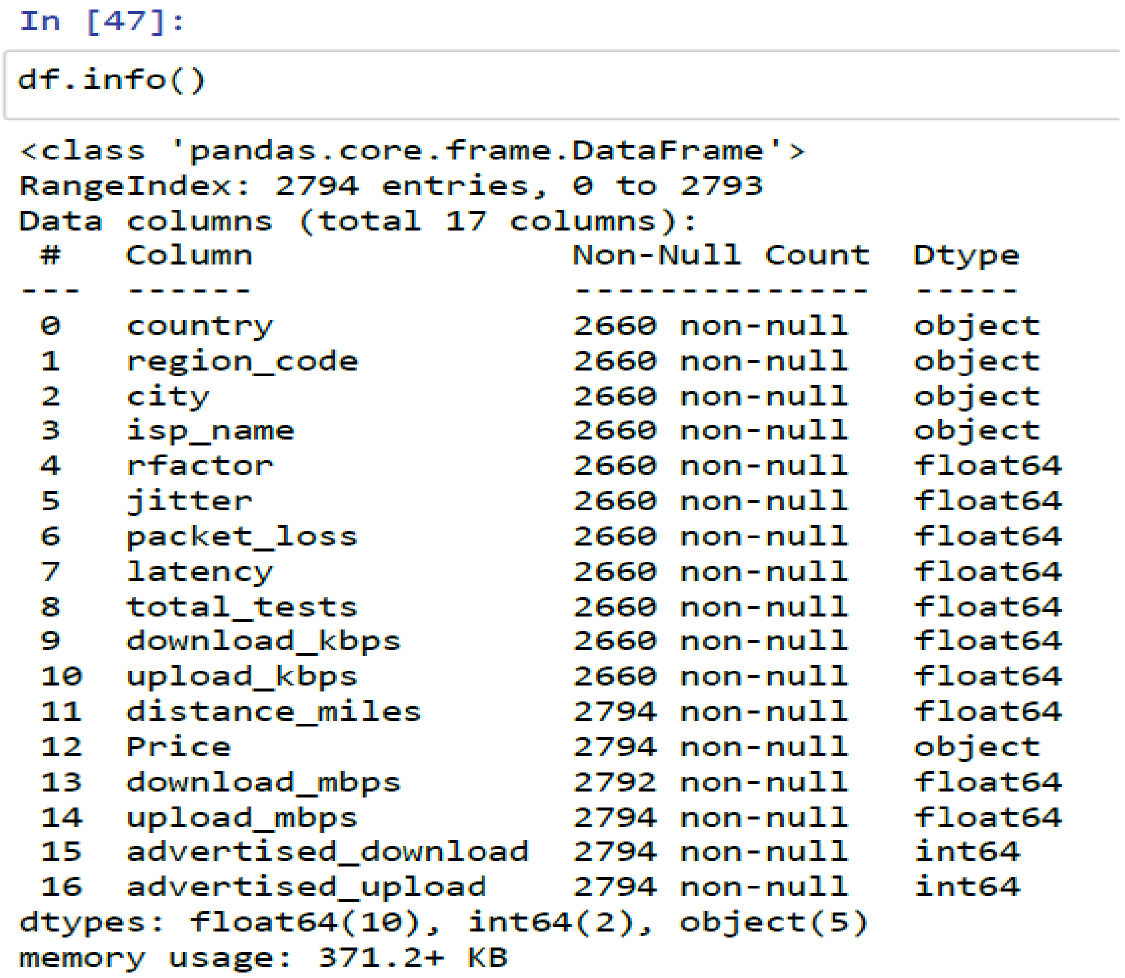
3. **Data Extraction and Attributes:** Data will be provided by CW-E and appropriate school boards. CW-E has provided us with a sample dataset based on CIRA speed test until we have access to the actual dataset from the respective school boards. In order to better understand the domain knowledge and various metrics that are supposed to be involved in the data, we have acquired datasets from multiple online sources like "Ookla Open Data Initiative" and "National Broadband Data Canada" and combined them into one dataset.

View of the dataset:



C	D	E	F	G	H	I	J	K	L
city	isp_name	rfactor	jitter	packet_loss	latency	total_tests	download_kbps	upload_kbps	distance_miles
Toronto	Bell Canada	86.3327	19.0047	1.74274	52.761	517	3365.82	744.312	232.303
Toronto	Bell Canada	85.7871	18.761	1.97105	52.2475	527	3350.43	737.829	231.689
Toronto	Bell Canada	85.9151	18.6485	1.92527	51.9407	541	3348.42	733.078	231.355
Toronto	Bell Canada	86.2588	18.1382	1.80647	51.0977	560	3361	722.224	230.957
Toronto	Bell Canada	86.2331	18.5027	1.80634	51.4178	566	3359.7	717.75	230.98
Toronto	Bell Canada	86.3507	18.0439	1.78246	50.0281	594	3348.82	715.784	231.761
Toronto	Bell Canada	86.4608	17.8983	1.74991	49.1832	619	3363.05	727.794	231.86
Toronto	Bell Canada	86.3697	17.5066	1.79973	48.6247	618	3361.14	736.59	231.834
Toronto	Bell Canada	86.436	17.1855	1.77568	49.0236	628	3360.79	737.993	232.152
Toronto	Bell Canada	86.3379	16.5379	1.8417	47.6434	643	3332.8	725.711	232.459
Toronto	Bell Canada	86.2154	16.4172	1.89601	47.3677	646	3337.83	736.31	232.578
Toronto	Bell Canada	86.2219	16.2412	1.8957	47.5007	647	3318.41	727.424	232.606
Toronto	Bell Canada	86.1934	16.427	1.90596	47.2524	629	3309.66	721.429	233.227
Toronto	Bell Canada	86.1823	16.3706	1.90986	47.4086	621	3320.05	735.683	233.283
Toronto	Bell Canada	86.1128	16.8152	1.92624	47.6789	615	3345.53	752.805	233.257
Toronto	Bell Canada	86.1043	16.8684	1.92889	47.6482	621	3328.12	741.186	233.328
Toronto	Bell Canada	86.2274	16.2953	1.89422	47.3392	621	3323.07	740.909	233.602
Toronto	Bell Canada	86.3044	16.4733	1.85843	47.4805	610	3313.06	736.297	233.621
Toronto	Bell Canada	86.3265	16.4774	1.84934	47.5016	613	3315.41	727.73	232.893
Toronto	Bell Canada	86.4373	15.6851	1.84747	44.8337	613	3316.57	722.339	233.124
Toronto	Bell Canada	86.1727	15.632	1.95523	44.7475	618	3313.54	724.472	233.232
Toronto	Bell Canada	86.0484	15.6686	2.00376	44.7937	621	3329.46	727.696	233.065
Toronto	Bell Canada	85.9117	15.9285	2.05198	44.9159	604	3338.57	733.039	233.197
Toronto	Bell Canada	86.2621	14.4629	1.95873	43.151	604	3357.11	736.559	233.178
Toronto	Bell Canada	86.2046	14.9578	1.95736	44.6047	605	3363.38	737.06	232.981
Toronto	Bell Canada	86.5782	13.906	1.84345	43.1581	564	3370.69	733.327	232.813
Toronto	Bell Canada	86.3645	13.3166	1.94569	42.6568	534	3369.24	730.784	233.131
Toronto	Bell Canada	86.7542	13.4598	1.78473	42.881	513	3372.31	736.143	233.38
Toronto	Bell Canada	86.7683	13.4044	1.78126	42.7725	514	3383.87	738.199	233.565
Toronto	Bell Canada	86.9234	13.8468	1.70574	43.2258	516	3399.27	747.022	233.874
Toronto	Bell Canada	86.7249	15.6979	1.72431	45.6147	501	3411.79	752.191	234.235
Toronto	Bell Canada	87.7402	20.3478	1.16052	51.9937	504	3440.99	759.636	234.016
Toronto	Bell Canada	87.7034	20.0466	1.17471	52.6408	514	3454.46	759.423	233.901
Toronto	Bell Canada	87.617	20.2149	1.20944	52.2903	507	3454.59	755.559	234.044

Attributes and its types:



```
In [47]:

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2794 entries, 0 to 2793
Data columns (total 17 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   country              2660 non-null   object 
 1   region_code          2660 non-null   object 
 2   city                 2660 non-null   object 
 3   isp_name             2660 non-null   object 
 4   rfactor              2660 non-null   float64
 5   jitter               2660 non-null   float64
 6   packet_loss          2660 non-null   float64
 7   latency              2660 non-null   float64
 8   total_tests          2660 non-null   float64
 9   download_kbps        2660 non-null   float64
 10  upload_kbps          2660 non-null   float64
 11  distance_miles       2794 non-null   float64
 12  Price                2794 non-null   object 
 13  download_mbps        2792 non-null   float64
 14  upload_mbps          2794 non-null   float64
 15  advertised_download  2794 non-null   int64  
 16  advertised_upload    2794 non-null   int64  
dtypes: float64(10), int64(2), object(5)
memory usage: 371.2+ KB
```

Definition of Attributes:

- **Jitter**: Jitter is when there is a time delay in the sending of data packets over a network connection
- **Packet Loss:** Packets are small units of data transmitted over a network from a particular source to a destination. Packet loss occurs when a network packet fails to reach its expected destination, resulting in information loss.
- **Latency:** Latency, also called ping, measures how much time it takes for your computer, the internet, and everything in between, to respond to an action you take (like clicking on a link).
- **ISP_name:** Names of internet service providers
- **Download and Upload kbps:** Download and Upload speed of internet.
- **Distance_miles:** Distance from the server to the end user.
- **Price:** Monthly price of user's broadband service
- **Total_test:** Average number of device tests.
- **Download and upload mbps:** Broadband speed in mbps
- **Advertised Download and Upload: Speed advertised by Internet Service Providers.**
- **City:** Name of the city
- **Country:** Name of the country.
- **Region:** Code of the region.
- **R-faction**: ratio of speed fluctuation.

4. Data Evaluation, Analysis and Improvement:

- **Creating Strong Data Processes**: This is a phase to implement processes in place for collecting, preparing, storing, and distributing the data.

- **Identifying the right tools** or platforms or technology solutions is essential to building a data management strategy.

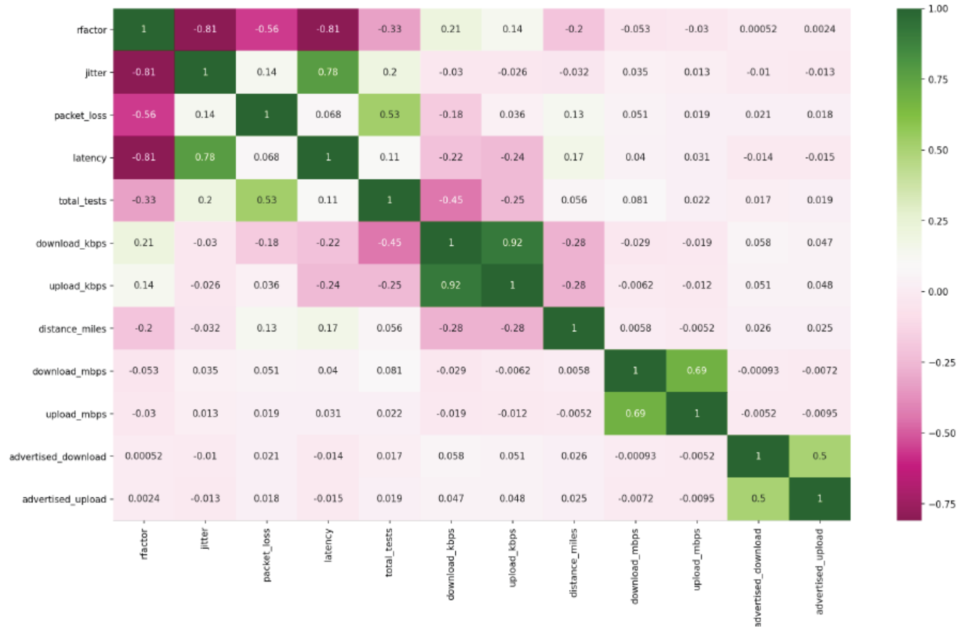
  Our planned tools include:

- **Data extraction** with support of **CW-E** and appropriate people at the **school board**.
- **Loading data** securely into SQL database by integrating **SSMS** (SQL Server Management Studio) with **Docker containers** or **Azure SQL Server**.
- **ETL** operations using **SSIS** (SQL Server Integration Services) .
- **Data Cleaning** and **Transformation** using **MS Excel** and **T-SQL querying**.
- **Analysing** broadband metrics such as data consumption, Latency and quality using python analysis which include **EDA** and **regression and tests**.
- Building **ML Prediction Model** for data consumption using **scikit-learn**.
- Building **dashboards** for **real-time monitoring** of broadband metrics using **Tableau** and **Azure Synapse Analytics.**

Analysis: To better understand the relationship between various variables we performed basic EDA Analysis using python and below are the results:

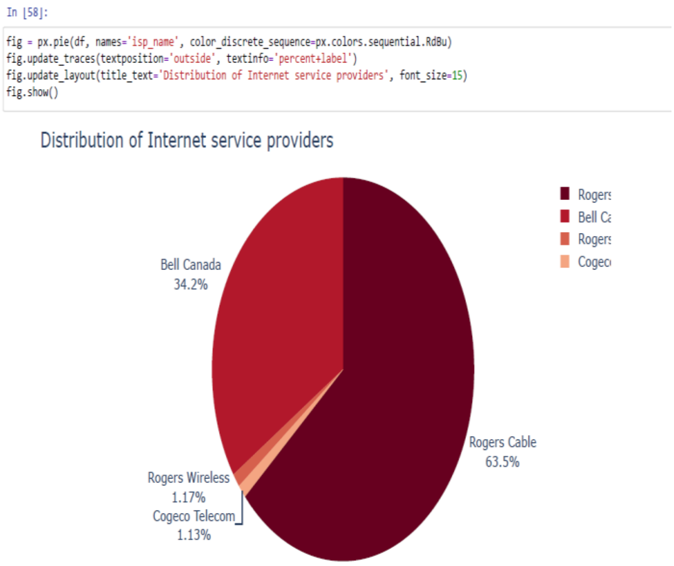
Correlation Plot:

```
<AxesSubplot:>
```



The above chart shows the correlation with different variables of the dataset like Jitter, Latency, Download and upload speed, Distance in miles.

Internet Service Provider’s Distribution:

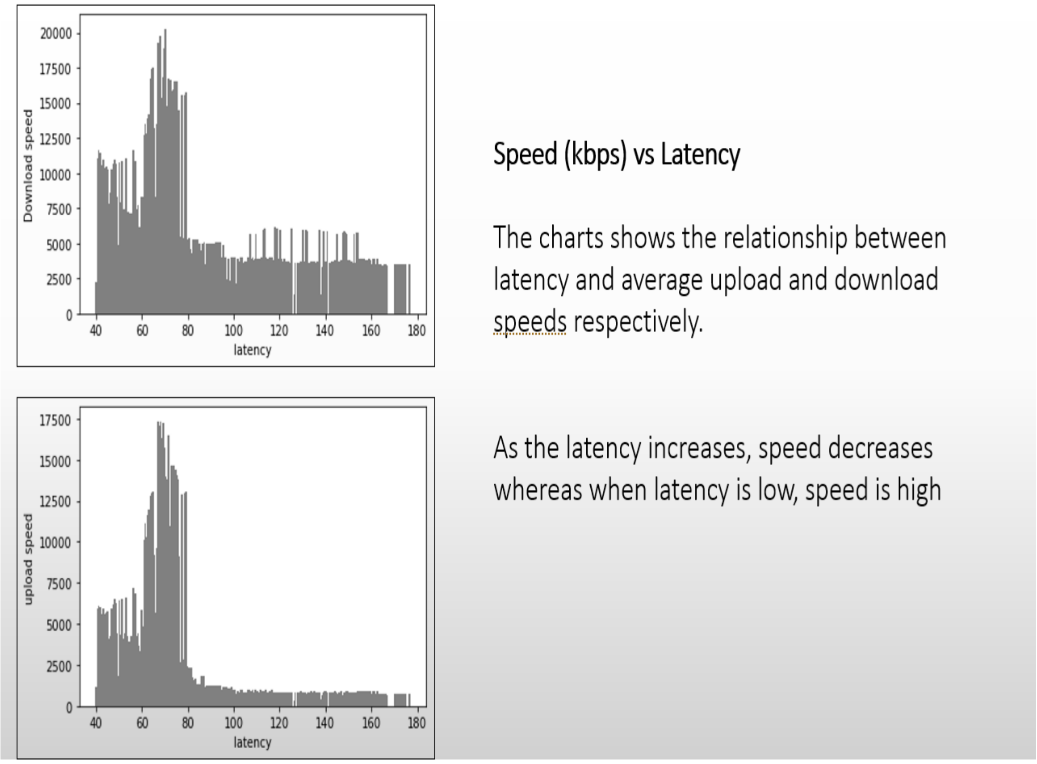


In [58]:

```
fig = px.pie(df, names='isp_name', color_discrete_sequence=px.colors.sequential.RdBu)
fig.update_traces(textposition='outside', textinfo='percent+label')
fig.update_layout(title_text='Distribution of Internet service providers', font_size=15)
fig.show()
```

According to the dataset, Rogers cable service is used by highest distribution in Toronto. In Windsor, Cogeco holds leading position.

Speed vs Latency



Speed (kbps) vs Latency

The charts shows the relationship between latency and average upload and download speeds respectively.

As the latency increases, speed decreases whereas when latency is low, speed is high

(Further analysis can be found in the python.pdf and PPT files attached with submission)

Data Improvement:

- Design and **develop improvement plans** based on **prior analysis which we performed using secondary data acquired from Ookla**. The plans should comprehend timeframes, resources.
- **Implement solutions** determined in the Improve stage. Comprehend both technical as well as any business process-related changes. Implement a comprehensive **'Change Management'** plan to **ensure that all stakeholders related to CW-E and school boards appropriately informed.**
- **Verify** at periodic intervals that the **data is consistent with the business goals and the data rules specified** in the definition step. **Communicate the Data Quality metrics and current status to all stakeholders** on a regular basis to ensure that Data Quality discipline is maintained on an ongoing basis across the organisation.

5. Data Security: Data security is defined as, the invested parties having correct protocols to access the project. It means that all the stakeholders including CW-E, School boards and St.Clair College involved should have the access to information and data according to their role.The encryption of the data is crucial in project management security and the data should not be available to everyone. Information and physical security both are important in this regard and we will be discussing both below.

1. **Educating our users**: The most vital step is **educating our group members and stakeholders relating to both CW-E and school boards** and. Everyone should be aware of the importance of the project. Changing our behavior towards how we interact with technology is highly important. A careless click on a phishing link can cost you (or your clients) a data breach along with significant financial damages.

2. **Use right security tools**: **Virtual Private Networks and I.P address and passwords relating SSMS databases make it easy for us to share data securely across the networks**. This crucial security tool enhances project protection, encrypts the data so that no third parties can decipher it, and makes communications for the group members and stakeholders protected.

3. **Backup** is extremely important in high stake projects. Our group leader and the stakeholder should store copies of data on external hard drives, Flash drives, and or on other devices in case of laptop theft, equipment damage, or any other unforeseen consequences. The copy of the data will help us go on with your strategic planning and will not compromise your deadlines and project itself. But we must also take great care of your external devices where we have stored data. Make it weatherproof and check it from time to time to make sure that the data is up to date and secure.

References:

Data Quality – Simple 6 Step Process – Digital Transformation for Professionals (digitaltransformationpro.com)

5 Key Steps to Creating a Data Management Strategy | Tableau

What Is Data Governance and Why Does It Matter? (techtarget.com)

Ookla's Open Data Initiative | Ookla®

Connecting Windsor-Essex (cw-e.ca)