QUANTUM WOLF

DATA INTELLIGENCE RESEARCH LAB

DATE: 14/03/2025 – 18/03/2025

NAME: SURIYA E

Predictive Corrosion Mapping in Underground Pipelines

1. Problem Statement

Underground pipelines, which transport water and gas, are critical infrastructure components. However, hidden corrosion beneath insulation is a major issue, causing **12% of leaks** and resulting in **$50 billion in repair costs annually**. Current manual inspection methods fail to detect **70% of early-stage corrosion**, leading to unexpected failures, environmental hazards, and significant financial losses. With **70% of U.S. pipelines over 50 years old**, there is an urgent need for a predictive system to identify corrosion hotspots early and prevent leaks.

2. Solution

The project aims to predict corrosion hotspots using **ground-penetrating radar (GPR)**, **soil sensors**, and **machine learning models**. The solution involves:

- Collecting and organizing data from GPR, soil pH, and leak history.
- Cleaning and preprocessing the data.
- Engineering features like corrosion rate and humidity index.
- Predicting corrosion using **multiple linear regression**.
- Performing risk analysis and mapping corrosion severity using **heatmaps**.
- Scheduling preventive maintenance using **Monte Carlo simulations**.
- Generating a **visual dashboard** for performance reporting.

3. Working Process of Underground Pipeline Corrosion Prediction

Step 1: Data Collection

- **Configuration**:
    - Open the Excel file **Corrosion_Prediction.xlsx.**
    - Create separate sheets for:
        - **GPR Data**: Columns for Location, GPR Signal Strength, Moisture Levels.
        - **Soil pH Data**: Columns for Location, Soil pH, Temperature.
        - **Leak History**: Columns for Location, Leak Incidents, Date.
        - **Weights**: Columns for Variables (Soil pH, Moisture, Temperature, GPR Signal) and their assigned weights.
- **Purpose**:
    - Organize raw data into structured formats for easy access and analysis.
- **Expected Output**:
    - A well-organized dataset with separate sheets for each data type, ready for cleaning and analysis.

Step 2: Data Cleaning

- **Configuration**:
    - Select the dataset (e.g., Soil pH Data).
    - Go to **Data → Remove Duplicates**.
    - For missing values, use:

      =IFERROR(A2, AVERAGE(A:A))

    - For outliers, use Z-Score or IQR methods (optional).

- **Purpose**:
    - Ensure data accuracy by removing duplicates, handling missing values, and managing outliers.
- **Expected Output**:
    - Clean datasets free from duplicates, missing values, and outliers, ensuring reliable analysis.

	A	B	C	D	E
1	Location	Soil_pH	Moisture	Temperature	GPR_Signal
2	Location_1	6.62	17.41	21.54	0.71
3	Location_2	8.35	31.68	21.17	0.82
4	Location_3	7.70	44.92	37.66	0.33
5	Location_4	7.30	39.29	21.24	0.66
6	Location_5	5.97	42.26	21.80	0.61
7	Location_6	5.97	36.35	33.98	0.85
8	Location_7	5.67	37.69	26.24	0.92
9	Location_8	8.10	43.97	34.42	0.11
10	Location_9	7.30	19.99	16.63	0.71
11	Location_10	7.62	29.58	27.19	0.15
12	Location_11	5.56	18.85	15.84	0.59
13	Location_12	8.41	49.51	16.57	0.36
14	Location_13	8.00	47.76	37.66	0.38
15	Location_14	6.14	11.58	18.48	0.42
16	Location_15	6.05	38.22	28.31	0.66
17	Location_16	6.05	47.01	25.28	0.40
18	Location_17	6.41	17.22	23.68	0.76
19	Location_18	7.07	32.72	37.50	0.46
20	Location_19	6.80	46.62	15.55	0.16
21	Location_20	6.37	11.36	31.59	0.81
22	Location_21	7.34	37.90	39.08	0.36
23	Location_22	5.92	21.89	29.00	0.49
24	Location_23	6.38	46.98	38.42	0.72
25	Location_24	6.60	48.84	16.31	0.40

Step 3: Feature Engineering

Add New Columns for Feature Engineering

- **Configuration**:
    - Add 3 new columns in the **GPR Data Sheet**:
        - **Column F**: Corrosion Risk Factor
        - **Column G**: Normalized GPR Signal
        - **Column H**: Severity Score
- **Purpose**:
    - Create new features to improve the predictive model's accuracy and provide deeper insights into corrosion risk.

1. Corrosion Risk Factor (F2)

- **Configuration**:
  - In Cell F2, input the formula: **=IF(AND(B3<6.5, C3>30, D3>35, E3<0.3), "High Risk", IF(AND(B3<7, C3>25, D3>30, E3<0.5), "Moderate Risk", "Low Risk"))**
  - Drag the fill handle down to apply the formula to all rows.
- **Purpose**:
  - Classify locations as **High Risk** , **Moderate Risk** and **Low Risk** based on soil pH (<6.5), moisture (>30), temperature (>35), and GPR signal (<0.3).

2. Normalized GPR Signal (G2)

- **Configuration**:
  - In Cell G2, enter the formula:

    =(E2-MIN(E:E))/(MAX(E:E)-MIN(E:E))

  - Drag the fill handle down to apply the formula to all rows.
- **Purpose**:
  - Normalize the GPR signal values to a range between 0 and 1 for better comparison and analysis.

3. Severity Score (H2)

- **Configuration**:
  - In Cell H2, enter the formula:

    =C2*D2*E2

  - Drag the fill handle down to apply the formula to all rows.
- **Purpose**:
  - Calculate a **Severity Score** by multiplying moisture, temperature, and GPR signal values to quantify corrosion severity.

Expected Outputs

1. **Corrosion Risk Factor**:
   - A column classifying locations as **High Risk** or **Low Risk** based on environmental and GPR signal conditions.
2. **Normalized GPR Signal**:
   - A column with GPR signal values normalized to a 0-1 range for easier comparison.
3. **Severity Score**:
   - A column with calculated Severity Scores, quantifying corrosion risk for each location.

	A	B	C	D	E	F	G	H
1	Location	Soil_pH	Moisture	Temperature	GPR_Signal	Corrosion Risk Factor	Normalized GPR Signal	Corrosion Severity Score
2	Location_1	6.62	17.41	21.54	0.71	Low Risk	0.672786768	264.51
3	Location_2	8.35	31.68	21.17	0.82	Moderate Risk	0.796901161	547.99
4	Location_3	7.70	44.92	37.66	0.33	Low Risk	0.250088527	550.43
5	Location_4	7.30	39.29	21.24	0.66	Low Risk	0.62490541	552.73
6	Location_5	5.97	42.26	21.80	0.61	Low Risk	0.571719018	566.19
7	Location_6	5.97	36.35	33.98	0.85	Low Risk	0.833089792	1049.53
8	Location_7	5.67	37.69	26.24	0.92	Moderate Risk	0.90642683	905.54
9	Location_8	8.10	43.97	34.42	0.11	Low Risk	0.011515999	167.88
10	Location_9	7.30	19.99	16.63	0.71	Low Risk	0.674105137	234.92
11	Location_10	7.62	29.58	27.19	0.15	Low Risk	0.05123855	117.93
12	Location_11	5.56	18.85	15.84	0.59	Low Risk	0.548806595	177.34
13	Location_12	8.41	49.51	16.57	0.36	Moderate Risk	0.287294123	294.32
14	Location_13	8.00	47.76	37.66	0.38	Low Risk	0.306458992	676.52
15	Location_14	6.14	11.58	18.48	0.42	Low Risk	0.352691552	89.36
16	Location_15	6.05	38.22	28.31	0.66	Low Risk	0.621319831	713.29
17	Location_16	6.05	47.01	25.28	0.40	Low Risk	0.333762274	476.08
18	Location_17	6.41	17.22	23.68	0.76	Moderate Risk	0.732848633	309.77
19	Location_18	7.07	32.72	37.50	0.46	Low Risk	0.404317	569.32
20	Location_19	6.80	46.62	15.55	0.16	Low Risk	0.067774069	117.06

Step 4: Corrosion Prediction Using Multiple Linear Regression

- **Configuration**:
  - In the **Weights Sheet**, assign weights to variables:
    - Soil pH: 0.3
    - Moisture: 0.4
    - Temperature: 0.2
    - GPR Signal: 0.1

- In the **Corrosion Prediction Sheet**, calculate corrosion scores using:

=SUMPRODUCT(B2:E2, Weights!$B$2:$B$5)

- **Purpose**:
  - Predict corrosion likelihood using a weighted linear regression model.
- **Expected Output**:
  - A column of corrosion scores for each location, indicating the likelihood of corrosion.

Step 5: Risk Analysis (Anomaly Detection)

- **Configuration**:
  - Select the Corrosion Score column (F2:F100).
  - Go to **Conditional Formatting → Highlight Cells Rules → Greater Than...**
    - Set > 0.8 (High Risk in Red).
    - Set 0.5 - 0.8 (Moderate Risk in Yellow).
    - Set < 0.5 (Low Risk in Green).
- **Purpose**:
  - Identify and highlight high-risk corrosion areas for immediate attention.
- **Expected Output**:
  - A visually formatted column showing risk levels (High, Moderate, Low) for each location.

	A	B	C	D	E	F	G	H
1	Location	Soil_pH	Moisture	Temperature	GPR_Signal	Corrosion Risk Factor	Normalized GPR Signal	Corrosion Severity Score
2	Location_1	6.62	17.41	21.54	0.71	Low Risk	0.672786768	264.51
3	Location_2	8.35	31.68	21.17	0.82	Moderate Risk	0.796901161	547.99
4	Location_3	7.70	44.92	37.66	0.33	Low Risk	0.250088527	550.43
5	Location_4	7.30	39.29	21.24	0.66	Low Risk	0.62490541	552.73
6	Location_5	5.97	42.26	21.80	0.61	Low Risk	0.571719018	566.19
7	Location_6	5.97	36.35	33.98	0.85	Low Risk	0.833089792	1049.53
8	Location_7	5.67	37.69	26.24	0.92	Moderate Risk	0.90642683	905.54
9	Location_8	8.10	43.97	34.42	0.11	Low Risk	0.011515999	167.88
10	Location_9	7.30	19.99	16.63	0.71	Low Risk	0.674105137	234.92
11	Location_10	7.62	29.58	27.19	0.15	Low Risk	0.05123855	117.93
12	Location_11	5.56	18.85	15.84	0.59	Low Risk	0.548806595	177.34
13	Location_12	8.41	49.51	16.57	0.36	Moderate Risk	0.287294123	294.32
14	Location_13	8.00	47.76	37.66	0.38	Low Risk	0.306458992	676.52
15	Location_14	6.14	11.58	18.48	0.42	Low Risk	0.352691552	89.36
16	Location_15	6.05	38.22	28.31	0.66	Low Risk	0.621319831	713.29
17	Location_16	6.05	47.01	25.28	0.40	Low Risk	0.333762274	476.08
18	Location_17	6.41	17.22	23.68	0.76	Moderate Risk	0.732848633	309.77
19	Location_18	7.07	32.72	37.50	0.46	Low Risk	0.404317	569.32
20	Location_19	6.80	46.62	15.55	0.16	Low Risk	0.067774069	117.06

Step 6: Corrosion Severity Mapping (Heatmap)

- **Configuration**:
    - Select the entire Corrosion Score Column.
    - Go to **Home → Conditional Formatting → Color Scale**.
    - Choose Red-Yellow-Green Scale.
- **Purpose**:
    - Visualize corrosion severity across all locations using a color-coded heatmap.
- **Expected Output**:
    - A heatmap showing corrosion severity, with high-risk areas in red and low-risk areas in green.

	A	B	C	D	E	F	G	H	I
1	Location	Soil_pH	Moisture	Temperature	GPR_Signal	Corrosion Risk Factor	Normalized GPR Signal	Corrosion Severity Score	Corrosion Severity Score
2	Location_1	6.62	17.41	21.54	0.71	Low Risk	0.672786768	264.51	264.51
3	Location_2	8.35	31.68	21.17	0.82	Moderate Risk	0.796901161	547.99	547.99
4	Location_3	7.70	44.92	37.66	0.33	Low Risk	0.250088527	550.43	550.43
5	Location_4	7.30	39.29	21.24	0.66	Low Risk	0.62490541	552.73	552.73
6	Location_5	5.97	42.26	21.80	0.61	Low Risk	0.571719018	566.19	566.19
7	Location_6	5.97	36.35	33.98	0.85	Low Risk	0.833089792	1049.53	1049.53
8	Location_7	5.67	37.69	26.24	0.92	Moderate Risk	0.90642683	905.54	905.54
9	Location_8	8.10	43.97	34.42	0.11	Low Risk	0.011515999	167.88	167.88
10	Location_9	7.30	19.99	16.63	0.71	Low Risk	0.674105137	234.92	234.92
11	Location_10	7.62	29.58	27.19	0.15	Low Risk	0.05123855	117.93	117.93
12	Location_11	5.56	18.85	15.84	0.59	Low Risk	0.548806595	177.34	177.34
13	Location_12	8.41	49.51	16.57	0.36	Moderate Risk	0.287294123	294.32	294.32
14	Location_13	8.00	47.76	37.66	0.38	Low Risk	0.306458992	676.52	676.52
15	Location_14	6.14	11.58	18.48	0.42	Low Risk	0.352691552	89.36	89.36
16	Location_15	6.05	38.22	28.31	0.66	Low Risk	0.621319831	713.29	713.29
17	Location_16	6.05	47.01	25.28	0.40	Low Risk	0.333762274	476.08	476.08
18	Location_17	6.41	17.22	23.68	0.76	Moderate Risk	0.732848633	309.77	309.77
19	Location_18	7.07	32.72	37.50	0.46	Low Risk	0.404317	569.32	569.32
20	Location_19	6.80	46.62	15.55	0.16	Low Risk	0.067774069	117.06	117.06

Step 7: Preventive Maintenance Scheduling Using Monte Carlo Simulation

- **Configuration**:
    - Create a new sheet named **Monte Carlo Simulation**.
    - In Column A, copy all Location Names (Location_1 to Location_20).
    - In Column B, copy the Corrosion Score from the previous calculation.
    - In Cell C2, enter:

=RAND() * B2

- Drag the formula down for 1000 rows for each location.
- In Column D, calculate the Failure Probability:

=COUNTIF(C2:C1001, ">=0.8") / 1000

- Visualize the Monte Carlo Simulation using a Bar Chart or Histogram.

- **Purpose**:
  - Simulate 1000 scenarios of corrosion failure to predict failure probabilities and schedule preventive maintenance.
- **Expected Output**:
  - A column of failure probabilities for each location and a visual chart showing failure patterns.

	C	D	E	F	G	H
1	Moisture	Temperature	GPR_Signal	Corrosion Severity Score	Simulation (1000 Runs)	Failure Probability
2	17.41	21.54	0.71	264.51	223.5	0.997
3	31.68	21.17	0.82	547.99	199.6	0.996
4	44.92	37.66	0.33	550.43	373.1	0.995
5	39.29	21.24	0.66	552.73	390.0	0.994
6	42.26	21.80	0.61	566.19	541.7	0.993
7	36.35	33.98	0.85	1049.53	608.6	0.992
8	37.69	26.24	0.92	905.54	604.5	0.991
9	43.97	34.42	0.11	167.88	18.9	0.99
10	19.99	16.63	0.71	234.92	84.7	0.989
11	29.58	27.19	0.15	117.93	80.2	0.988
12	18.85	15.84	0.59	177.34	31.7	0.987
13	49.51	16.57	0.36	294.32	233.7	0.986
14	47.76	37.66	0.38	676.52	281.0	0.985
15	11.58	18.48	0.42	89.36	83.3	0.984

Step 8: Performance Report (Graph Dashboard)

- **Configuration**:
  - **Pivot Table for Corrosion Severity**:
    - Select your Corrosion Prediction Sheet.
    - Click on **Insert → Pivot Table**.
    - Choose "New Worksheet" and click OK.
    - Configure the Pivot Table:

- Rows: Location
- Values: Corrosion Severity (set to "Count")

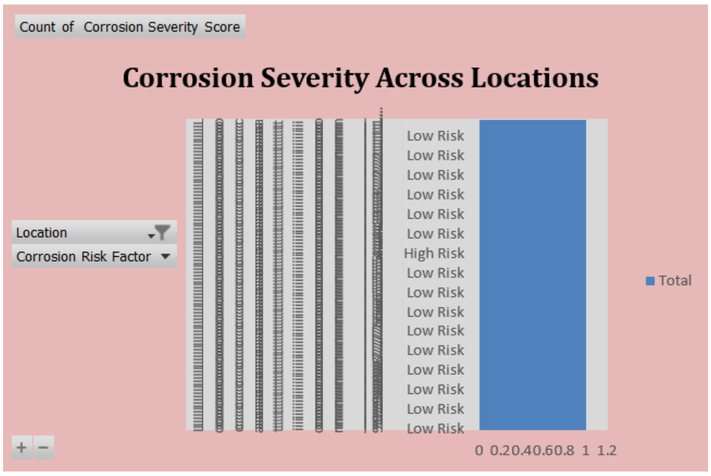
○ **Output:** A Pivot Table summarizing corrosion severity.

	A	B
3	Row Labels	Count of Corrosion Severity Score
4	⊟ **Location_999**	**1**
5	Low Risk	1
6	⊟ **Location_998**	**1**
7	Low Risk	1
8	⊟ **Location_997**	**1**
9	Low Risk	1
10	⊟ **Location_996**	**1**
11	Moderate Risk	1
12	⊟ **Location_995**	**1**
13	Low Risk	1
14	⊟ **Location_994**	**1**
15	Low Risk	1
16	⊟ **Location_993**	**1**
17	Low Risk	1
18	⊟ **Location_992**	**1**
19	Low Risk	1
20	⊟ **Location_991**	**1**
21	Low Risk	1
22	⊟ **Location_990**	**1**
23	Low Risk	1
24	⊟ **Location_99**	**1**
25	Moderate Risk	1
26	⊟ **Location_989**	**1**
27	Moderate Risk	1
28	⊟ **Location_988**	**1**

○ **Bar Chart for High-Risk Locations**:

- Select the Pivot Table Data.
- Click **Insert → Bar Chart**.
- Format the chart:
  - High-Risk locations in Red.
  - Moderate-Risk locations in Orange.
  - Low-Risk locations in Green.
- Add Title: "Corrosion Severity Across Locations".

○ **Output:** A Bar Chart showing high-risk locations.

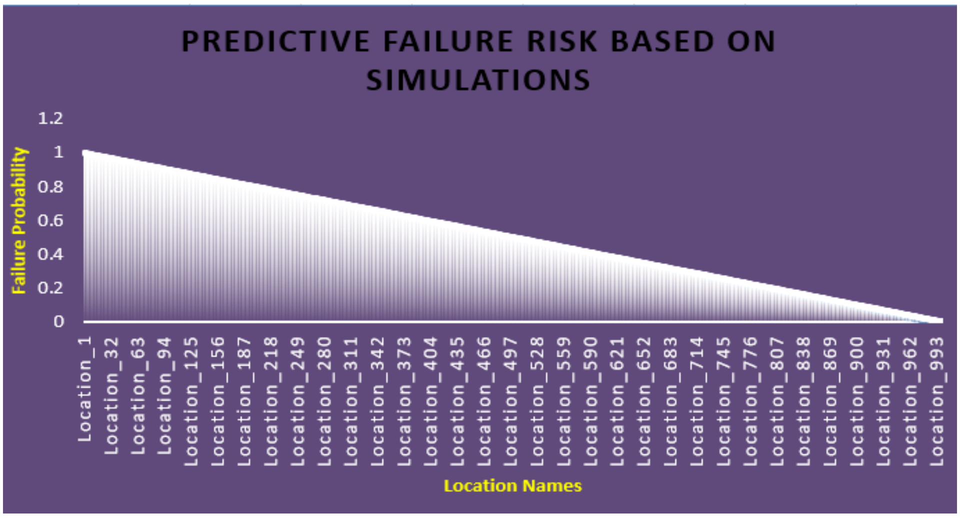


- **Heatmap for Corrosion Mapping**:
    - Select the Corrosion Severity Column.
    - Click **Home → Conditional Formatting → Color Scales**.
    - Choose Red-Yellow-Green Scale.
- **Output:** A Heatmap for corrosion mapping.

	A	B	C	D	E	F
1	Location	Soil_pH	Moisture	Temperature	GPR_Signal	Corrosion Severity Score
2	Location_1	6.62	17.41	21.54	0.71	264.51
3	Location_2	8.35	31.68	21.17	0.82	547.99
4	Location_3	7.70	44.92	37.66	0.33	550.43
5	Location_4	7.30	39.29	21.24	0.66	552.73
6	Location_5	5.97	42.26	21.80	0.61	566.19
7	Location_6	5.97	36.35	33.98	0.85	1049.53
8	Location_7	5.67	37.69	26.24	0.92	905.54
9	Location_8	8.10	43.97	34.42	0.11	167.88
10	Location_9	7.30	19.99	16.63	0.71	234.92
11	Location_10	7.62	29.58	27.19	0.15	117.93
12	Location_11	5.56	18.85	15.84	0.59	177.34
13	Location_12	8.41	49.51	16.57	0.36	294.32
14	Location_13	8.00	47.76	37.66	0.38	676.52
15	Location_14	6.14	11.58	18.48	0.42	89.36
16	Location_15	6.05	38.22	28.31	0.66	713.29
17	Location_16	6.05	47.01	25.28	0.40	476.08
18	Location_17	6.41	17.22	23.68	0.76	309.77
19	Location_18	7.07	32.72	37.50	0.46	569.32
20	Location_19	6.80	46.62	15.55	0.16	117.06

- **Predictive Failure Chart (Monte Carlo Data)**:
    - Go to Monte Carlo Simulation Sheet.
    - Select Failure Probability Data.
    - Click **Insert → Line Chart**.
    - Customize:
        - X-axis = Location Names

    - Y-axis = Failure Probability
    - Add Title: "Predictive Failure Risk Based on Simulations".
  - **Output:** A Predictive Failure Chart based on Monte Carlo simulations.



4. Final Outcomes and Workflow

- **Final Outputs**

  Outcome 1: Identified **high-risk corrosion hotspots** using predictive modeling.
  - A list of locations classified as High, Moderate, or Low Risk

  Outcome 2: Generated a heatmap for visualizing **corrosion severity** across pipeline locations.
  - A color-coded map showing corrosion severity across all locations.

  Outcome 3: Developed a **preventive maintenance schedule** using Monte Carlo simulations.
  - A schedule with failure probabilities for each location.

  Outcome 4: Created a **visual dashboard** for real-time monitoring and reporting.
  - A comprehensive dashboard for real-time monitoring and reporting.

Workflow:

1. Data Collection → 2. Data Cleaning → 3. Feature Engineering → 4. Corrosion Prediction → 5. Risk Analysis → 6. Severity Mapping → 7. Maintenance Scheduling → 8. Performance Reporting.

5. Problems Faced During the Project

1. **Heatmap Visualization**:
   - Difficulty in scaling the heatmap for large datasets.
   - Addressed by using pivot tables and conditional formatting.
2. **Monte Carlo Simulation**:
   - High computational load for 1000 simulations.
   - Optimized by using Excel's **Data Table** feature.

6. Conclusion

The **Predictive Corrosion Mapping in Underground Pipelines** project successfully addresses the critical issue of hidden corrosion in aging pipelines, which leads to leaks, environmental hazards, and significant financial losses. By leveraging **Excel-based tools**, **predictive modeling**, and **data visualization techniques**, the project provides a cost-effective and scalable solution for early detection and prevention of pipeline corrosion.

Impact:

- **Cost Savings**: Reduces annual repair costs by preventing leaks and addressing corrosion early.
- **Improved Safety**: Minimizes environmental and safety risks associated with pipeline failures.
- **Extended Lifespan**: Enhances the longevity of aging pipelines, 70% of which are over 50 years old in the U.S.
- **Resource Optimization**: Allocates maintenance resources efficiently, focusing on high-risk areas.