**Munich Re Assessment**

**Objective**

XYZ is one of the leading insurer that has identified a need to set up an automated pipeline for its quarterly data processing and consolidation activity.

The team receives the policy file quarterly, which they load into a consolidated table used for further analysis and reporting. As these are critical reports, the file has to undergo stringent data quality checks to ensure that the results are reliable and accurate. Currently, the entire process requires manual intervention.

You are provided with below data files

1. Datafile\_Test1.xlsx: Consolidated data till 2020 Q2



1. Datafile\_Test2.xlsx: Quarterly data file with issuances in 2021 Q1



You are required to

1. Prepare a script that
   1. Performs data validations on the input file (Datafile\_Test2.xlsx). Key areas to consider are data completeness, validity, accuracy, consistency, integrity.
   2. Creates a table to identify difference in information for the common contracts in file “Datafile\_Test1.xlsx” and “Datafile\_Test2.csv”. The table should provide for each common contract,

* the parameter for which there is a value change
* provide the value appearing in the 2 datasets for that parameter

For example, if Contract\_Number “A” appears in the 2 files with a difference in the value of age and gender, the table should give

|  |  |  |  |
| --- | --- | --- | --- |
| A | Age | 20 | 24 |
| A | Gender | M | F |

* 1. Adds a new column New/Old based on If Policy Date >='15-Feb-2021 =New Else Old rule.
  2. Appends “Datafile\_Test2” to “Datafile\_Test1” to create an exhaustive, standardized dataset that is required to be used for reporting business statistics. Generate a report with following information on the consolidated dataset-
* Range of Policy Date
* Number of policies in each month for 2020-2021
* Distribution of business by Age, Gender and Channel

1. Create a data quality scorecard for the input file. This will be used by the end user to gain an overview on data quality and assessing if data is fit for deriving insights. You may also include charts/graphs/other visuals.
2. Define an ETL workflow/process flow that automates and streamlines the processes in point 1 and 2, ensuring that a scalable, trusted and accurate data is available in real time.

Share the **scripts, reports/documentation.**