Data Analysis for North American Stainless

Presented By:

DSC-200-Group1

Aditya Khanal

Gaurab Baral

Mandeep Aryal

Shiva Khatri

Introduction (Background /Overview):

This report will briefly discuss how we systematically handled the cleaning and pre-processing of the dataset provided by North American Steels (NAS) and highlight our approach and findings obtained after the Data Analysis. Firstly, let us get to know more about NAS. North American Stainless is recognized as one of the world's largest stainless-steel producers; NAS leverages its access to worldwide resources and established foothold in key international markets to cater to customers beyond the borders of the United States. Their most significant asset, however, is their data in making data-driven decisions and improving business strategies. In this project, we implemented data extraction, data cleaning, and data visualization on the dataset using Python libraries like NumPy, Pandas, Matplotlib, and Seaborn to achieve valuable insights.

Our team initially organized and studied the dataset carefully. Next, we divided the dataset between each team member and analyzed it accordingly. Every member's objective was to extract meaningful facts from these datasets, and this report is based on the information retrieved from analyzing these datasets.

Problem Description:

1. Collection of production flaws, etc., observed from an automated system that identifies flaws in production.
2. Collection of production flaws, etc., observed from employees overseeing/managing the production lines.
3. Collection of production flaws, etc., observed from employees overseeing/managing the production lines.

Dataset Description:

- **AP4_Ptec_Coils:**

This csv file contains a list of coils processed by the automated inspection system (AVIS) at AP4 process line, provides a link between the system ID and the Mill product number. The AP4_Ptec coils file contains 2944 rows and 21 columns with "CoilId" as the primary key. The dataset has unique values of "CoilID" from the Id 467650 to the Id 470593. The columns in this data frame are:

```
Index(['CoilId', 'StartTime', 'EndTime', 'ParamSet', 'Grade', 'Length',
       'Width', 'Thickness', 'Weight', 'Charge', 'MaterialId', 'Status',
       'BdeCoilId', 'Description', 'LastDefectId', 'TargetQuality',
       'PdiRecvTime', 'SLength', 'InternalStatus', 'DefectCount', 'Campaign'],
      dtype='object')
```

- **AP4_Defects_Maps_10_coils:**

This csv file contains mapped defect records from the AP4 AVIS, limited to 10 coils. The file contains 35842 rows and 31 columns with "CoilId" as the column that links with the AP4_Ptec_coil file. The columns in this dataset are:

```
Index(['CoilId', 'DefectId', 'Class', 'Grade', 'PeriodId', 'PeriodLength',
       'PositionCD', 'PositionRCD', 'PositionMD', 'Side', 'SizeCD', 'SizeMD',
       'CameraNo', 'DefectNo', 'MergedTo', 'Confidence', 'RoiX0', 'RoiX1',
       'RoiY0', 'RoiY1', 'OriginalClass', 'PP_ID', 'PostCL', 'MergerPP',
       'OnlineCPP', 'OfflineCPP', 'Rollerid', 'InternalStatus',
       'CL_PROD_CLASS', 'CL_TEST_CLASS', 'AbsPosCD'],
      dtype='object')
```

- **claims_2023-05:**

The dataset contains customer claims for the month of June 2023. This file has had the unnecessary and extremely sensitive data removed already, primarily customers and internal personnel IDs. There are 100 rows and 66 columns in this dataset. The columns in the dataset are:

```
Index(['ClaimSource', 'ClaimNumber', 'ClaimDispositionSequence',
       'BusinessUnit', 'ClaimType', 'ProductIdentification1',
       'ProductIdentification2', 'MaterialSource', 'Heat', 'CastDate',
       'ProductType', 'SteelFamily', 'SteelType', 'SteelGradeASTMAISI',
       'Finish', 'MaterialGaugeOrDiameter', 'MaterialWidthOrLegLength',
       'MaterialLength', 'CustomerNumber', 'CustomerDestination',
       'OrderAbbreviation', 'OrderNumber', 'OrderItem', 'NationalExportCode',
       'SisterDivisionsClaimed', 'Format', 'FormatMinGaugeOrDiameter',
       'FormatMaxGaugeOrDiameter', 'FormatMinWidthOrLegLength',
       'FormatMaxWidthOrLegLength', 'FormatMinWeight', 'FormatMaxWeight',
       'InvoiceSeries', 'InvoiceYear', 'InvoiceNumber', 'InvoiceItem',
       'OriginalShippedWeight', 'OriginalShipQuality',
       'ClaimDispositionStatus', 'ClaimCreateDate', 'QCApprovedDate',
       'ClosedDate', 'TotalWeightClaimed', 'CustomerClaimDefect',
       'CustomerClaimDefectDesc', 'CustomerClaimDefectWeight',
       'NASIdentifiedDefect', 'NASIdentifiedDefectDesc',
       'NASIdentifiedDefectWeight', 'AreaofResponsibilityDefect',
       'AreaofResponsibilityDefectDesc', 'AreaofResponsibilityDefectWeigh',
       'CustomerDefectOrigin', 'CustomerDefectGroup',
       'CustomerDefectGroupDesc', 'TotalReturnInventoryWeight',
       'TotalScrapAtCustomerWeight', 'TotalSell3rdPartyWeight',
       'TotalCustomerCreditWeight', 'LastInspectionLine',
       'LastInspectionMachine', 'LastInspectedDate', 'GeneralComment1',
       'GeneralComment2', 'GeneralComment3', 'GeneralComment4'],
      dtype='object')
```

- **TblFLInspection:**

This dataset contains the total inspection parameters record focused on product characterizations. This dataset contains 5000 rows and 77 columns. The column “FLInspectionID” is the primary key in this dataset. The columns in this dataset are:

```
Index(['FLInspectionID', 'LineID', 'InspectionDate', 'InspectionDateInt',
       'InspectionTime', 'InspectionTimeInt', 'DealerCode',
       'SuperiorFinishCode', 'InspectionNumber', 'ExitCoilNumber',
       'ExitCoilDivision', 'PackProductCode', 'SteelGradeID', 'CurrentGuage',
       'HotAPGuage', 'ColdAPGuage', 'CurrentWidth', 'InitialWidth',
       'NetWeight', 'TotalLength', 'TotalSheets', 'Percent1AQualityExt',
       'Percent1BQualityExt', 'Percent2QualityExt', 'PercentScrapQualityExt',
       'Percent1AQualityIntCAP', 'Percent1BQualityIntCAP',
       'Percent2QualityIntCAP', 'PercentScrapQualityIntCAP',
       'Percent1AQualityExtCAP', 'Percent1BQualityExtCAP',
       'Percent2QualityExtCAP', 'PercentScrapQualityExtCAP',
       'Percent1AQualityIntHAP', 'Percent1BQualityIntHAP',
       'Percent2QualityIntHAP', 'PercentScrapQualityIntHAP',
       'Percent1AQualityExtHAP', 'Percent1BQualityExtHAP',
       'Percent2QualityExtHAP', 'PercentScrapQualityExtHAP', 'MnDefect1',
       'DefectGroup1', 'DefectGroup2', 'MnDefect2', 'MnDefectCAPInt1',
       'MnDefectCAPInt2', 'MnDefectCAPExt1', 'MnDefectCAPExt2',
       'CAPDefectiveLength', 'MnDefectHAPInt1', 'MnDefectHAPInt2',
       'MnDefectHAPExt1', 'MnDefectHAPExt2', 'CAPSolution', 'HAPSolution',
       'CutLineSolution', 'ChangeOfSide', 'CAPLineGrpCode', 'HAPLineGrpCode',
       'ZMillLineGrpCode', 'CutLineGrpCode', 'CAPWorkCode', 'HAPWorkCode',
       'CutLineWorkCode', 'CAPYield', 'HAPYield', 'CutLineYield',
       'AccumulatedEfficiency', 'CreateProgram', 'CreateDate', 'CreateTime',
       'ChangeProgram', 'ChangeDate', 'ChangeTime', 'isActive',
       'InspectionDateTime'],
      dtype='object')
```

- **FLInspectionComments:**

This dataset contains comments from inspectors on product characterization and quality, links to product ID via tblFLInspection. The dataset contains 5000 rows and 12 columns. The column "FLInspectionCommentID" is the primary key however the column "FLInspectionID" merges with the previous dataset. The columns of this dataset are:

```
Index(['FLInspectionCommentID', 'FLInspectionID', 'DefectMapSeqNumber',
       'DefectMapRemarkSeqNumber', 'Comment', 'CreateProgram', 'CreateDate',
       'CreateTime', 'ChangeProgram', 'ChangeDate', 'ChangeTime', 'isActive'],
      dtype='object')
```

- **TblFlatInspectionProcesses:**

This dataset contains the per-process inspection status records. There are 5000 rows and 75 columns. The column "InspectionProcessID" is the primary key of this csv file. The list of columns in this dataset are:

```
Index(['InspectionProcessID', 'FlatCoilID', 'CoilNumber', 'LineID',
       'ProcessStartTime', 'InspectionStartTime', 'InspectionEndTime',
       'ApprovedTime', 'InspectionGroup', 'InspectionStatusID',
       'LateralEdgeSeamTopOS', 'LateralEdgeSeamTopMS',
       'LateralEdgeSeamBottomOS', 'LateralEdgeSeamBottomMS', 'InspectionType',
       'Observations', 'BuffTopHead', 'BuffTopCenter', 'BuffTopTail',
       'BuffBottomHead', 'BuffBottomCenter', 'BuffBottomTail', 'C47HeadHeight',
       'C47MiddleHeight', 'C47TailHeight', 'HeadPitch', 'MiddlePitch',
       'TailPitch', 'C09HeadHeight', 'C09MiddleHeight', 'C09TailHeight',
       'RoughnessTHeadOSSeverity', 'RoughnessTHeadCenterSeverity',
       'RoughnessTHeadDSSeverity', 'RoughnessTBodyOSSeverity',
       'RoughnessTBodyCenterSeverity', 'RoughnessTBodyDSSeverity',
       'RoughnessTTailOSSeverity', 'RoughnessTTailCenterSeverity',
       'RoughnessTTailDSSeverity', 'RoughnessBHeadOSSeverity',
       'RoughnessBHeadCenterSeverity', 'RoughnessBHeadDSSeverity',
       'RoughnessBBodyOSSeverity', 'RoughnessBBodyCenterSeverity',
       'RoughnessBBodyDSSeverity', 'RoughnessBTailOSSeverity',
       'RoughnessBTailCenterSeverity', 'RoughnessBTailDSSeverity',
       'RoughnessTHeadOSType', 'RoughnessTHeadCenterType',
       'RoughnessTHeadDSType', 'RoughnessTBodyOSType',
       'RoughnessTBodyCenterType', 'RoughnessTBodyDSType',
       'RoughnessTTailOSType', 'RoughnessTTailCenterType',
       'RoughnessTTailDSType', 'RoughnessBHeadOSType',
       'RoughnessBHeadCenterType', 'RoughnessBHeadDSType',
       'RoughnessBBodyOSType', 'RoughnessBBodyCenterType',
       'RoughnessBBodyDSType', 'RoughnessBTailOSType',
       'RoughnessBTailCenterType', 'RoughnessBTailDSType', 'HeadDefectCode',
       'TailScrap', 'HeadScrap', 'TailDefectCode', 'SamplesTaken', 'PaperUsed',
       'UserID', 'active'],
      dtype='object')
```

- **TblFlatInspectionMappedDefects:**
  This dataset contains manually mapped defects from each process line inspection, links to product ID via tblFlatInspectionProcesses. The dataset contains 5000 rows and 14 columns out of which the column “InspectionProcessID” serves as a primary key and can be combined for further analysis.

```
Index(['InspectionMappedDefectID', 'InspectionProcessID', 'DefectCodeID',
       'SideID', 'FaceID', 'StartPosition', 'Length', 'QualityID',
       'DefectCount', 'Description', 'FaceDescription', 'QualityCode',
       'QualityDescription', 'SideDescription'],
      dtype='object')
```

Dataset Samples:

- ## AP4_Ptec_Coils:

```
CoilId,StartTime,EndTime,ParamSet,Grade,Length,Width,Thickness,Weight,Charge,MaterialId,Status,BdeCoilId,Description,LastDefectId,TargetQuality,PdiRecvTime,SLength,Intern
467650,9/16/2023 10:20:20:46,9/16/2023 10:42:42:43,8,2,716034,1092,3.5,21320,,380,F,02AC2L,Inserted by TCP/IP BDE Server ,5998,-1,9/16/2023 9:51:51:36,724509,X,1458,91080
467651,9/16/2023 10:42:42:43,9/16/2023 11:08:08:42,5,-1,648440,936,3.5,17091,,61,F,03Y83H,Inserted by TCP/IP BDE Server ,9496,-1,9/16/2023 10:11:11:22,661111,X,3993,71761
467652,9/16/2023 11:08:08:42,9/16/2023 11:23:23:56,1,2,550688,1247,3.35,17849,,181,F,03X77T,Inserted by TCP/IP BDE Server ,3554,-1,9/16/2023 10:27:27:07,541324,X,2088,400
```

- ## AP4_Defects_Maps_10_coils:

```
1    CoilId,DefectId,Class,Grade,PeriodId,PeriodLength,PositionCD,PositionRCD,PositionMD,Side,SizeCD,SizeMD,CameraNo,DefectNo,MergedTo,Confidence,
2    467740,6321,267,5,14,6812,371,816,92931,0,59,102250,13,160,-2,0,242,258,11,46,-1,1000,282,-1,-1,9999,NULL,X,282,0,NULL
3    467740,6320,267,5,13,6302,406,655,44814,0,188,174164,13,37,-2,0,573,672,9,87,-1,1000,0,-1,-1,9999,NULL,X,0,0,NULL
4    467740,6319,267,5,12,6351,383,686,93172,0,178,125645,13,162,-2,0,216,276,5,120,-1,1000,259,-1,-1,9999,NULL,X,259,259,NULL
```

- ## claims_2023-05:

```
ClaimSource,ClaimNumber,ClaimDispositionSequence,BusinessUnit,ClaimType,ProductIdentification1,ProductIdentification2,MaterialSource,Heat,CastDate,ProductType,SteelFamily,SteelType,SteelGradeASTMAISI
S,69763,1,F,T,01X83T ,C     ,F,X83T,4/5/2023,B ,B,150,304L          ,22,0.0575,60,0,1005,3,AN,1012313,10,N,N,27,0.057,0.063,60,60.125,15000,25000,N ,2023,10,0,15550,1A,C,5/23/2023,6/22/2023,6/27/2023
S,69747,1,X,T,02X20Y ,BA    ,F,X20Y,1/27/2023,B ,A,101,301,12,0.0276,0,0,8030,1,PN,135508,2,N,N,0,0,0,0,0,0,0,BX,2023,0,0,5780,1,C,5/22/2023,6/26/2023,6/27/2023,1254,H43,Veiny Acid Stains
S,69410,1,F,T,03X44W ,C     ,F,X44W,2/20/2023,B ,C,270,316L          ,22,0.046,48,0,305,31,AN,1007622,1,N,N,417,0.045,0.047,48,48.0625,12000,20000,N ,2023,1,0,17940,5A,C,5/2/2023,6/1/2023,6/2/2023,12
```

- ## TblFLInspection:

```
FLInspectionID,LineID,InspectionDate,InspectionDateInt,InspectionTime,InspectionTimeInt,DealerCode,SuperiorFinishCode,InspectionNumber,ExitCoilNumber,ExitCoilDivision,PackProductCode,SteelGradeID,CurrentGuage,HotAPGuage,Col
3250536,30,9/1/2023 0:00,1230901,1/1/1900 20:23,202254,5000,32,4372092,05Y98H,     ,B,3,0.7112,4.0005,0.7112,1252.22,1252.22,24493.98798,3200400,0,6.63,0.03,93.34,0,0,99.69,0.29,0,0,99.69,0.29,0.02,100,0,0,0,100,0,0,0,212,
3250536,30,9/1/2023 0:00,1230901,1/1/1900 20:23,202254,5000,32,4372092,05Y98H,     ,B,3,0.7112,4.0005,0.7112,1252.22,1252.22,24493.98798,3200400,0,6.63,0.03,93.34,0,0,99.69,0.29,0,0,99.69,0.29,0.02,100,0,0,0,100,0,0,0,212,
3250536,30,9/1/2023 0:00,1230901,1/1/1900 20:23,202254,5000,32,4372092,05Y98H,     ,B,3,0.7112,4.0005,0.7112,1252.22,1252.22,24493.98798,3200400,0,6.63,0.03,93.34,0,0,99.69,0.29,0,0,99.69,0.29,0.02,100,0,0,0,100,0,0,0,212,
```

- ## FLInspectionComments:

```
FLInspectionCommentID,FLInspectionID,DefectMapSeqNumber,DefectMapRemarkSeqNumber,Comment,CreateProgram,CreateDate,CreateTime,ChangeProgram,ChangeDate,ChangeTime,isActive
8169735,3250536,10,10,ok                                      ,L3MNEFR   ,1230831,192201,L3MNEFR   ,1230831,192201,1
8169736,3250536,20,10,ok                                      ,L3MNEFR   ,1230831,192201,L3MNEFR   ,1230831,192201,1
8169737,3250536,30,10,ok                                      ,L3MNEFR   ,1230831,192201,L3MNEFR   ,1230831,192201,1
8169738,3250536,40,10,light  send to skp2                     ,L3MNEFR   ,1230831,192201,L3MNEFR   ,1230831,192201,1
```

- ## TblFlatInspectionProcesses:

```
InspectionProcessID,FlatCoilID,CoilNumber,LineID,ProcessStartTime,InspectionStartTime,InspectionEndTime,ApprovedTime,InspectionGroup,InspectionStatusID,LateralEdgeSeamTopOS,LateralEdgeSeamTop
1007943,6468471,02AA8N     ,27,55:00.0,25:29.7,45:41.3,52:54.7,NULL,Approved  ,0,0,0,0,M,"EK/WC70/24""ID/ J44 IAXA LGT CLDY/L47 LGT STN MRK XA/H06 IAXA SCTD LGT SHLOW SEE @ ANGLE/  980 NO TR
1007946,6471628,04AA8M     ,336,43:00.0,02:26.6,47:39.7,49:06.9,NULL,Approved  ,0,0,0,0,M,"MC/WC76/PPR/NO CORE/20""ID/CERT A/*SCHENK NOT WORKING*/HZY DLL APPRNC IAXA/ V45 IAXA- SN @ ANGLE N
1007947,6471628,04AA8M     ,336,43:00.0,02:26.6,47:39.7,49:06.9,NULL,Approved  ,0,0,0,0,M,"MC/WC76/PPR/NO CORE/20""ID/CERT A/*SCHENK NOT WORKING*/HZY DLL APPRNC IAXA/ V45 IAXA- SN @ ANGLE N
```

- ## TblFlatInspectionMappedDefects:

```
InspectionMappedDefectID,InspectionProcessID,DefectCodeID,SideID,FaceID,StartPosition,Length,QualityID,DefectCount,Description,FaceDescription,QualityCode,QualityDescription,Sid
6065043,1007942,263,4,3,1,12518,1,1,STAINS FROM ALKALI SECTION,Both,1,First,AW
6065044,1007942,290,4,2,1,12518,1,1,"BRIDLE ROLL MARK AP4-255""/1-152""/2-225""/3-99""",Exterior,1,First,AW
6065045,1007942,214,4,3,1,12518,1,1,DENTS FROM COLD ANNEALING & PICKLING LINE,Both,1,First,AW
6065046,1007942,58,4,3,1,12518,1,1,SLAB GRINDER GRAIN MARKS,Both,1,First,AW
6065047,1007942,296,2,2,2,12517,2,1,LINE SCRATCHES FROM HOT ANNEALING AND PICKLING,Exterior,2,Second,MS
6065048,1007942,310,3,2,3490,7110,1,1,WATER STAINS,Exterior,1,First,C
6065049,1007942,472,4,3,12518,1,2,1,Without Test Sample,Both,2,Second,AW
```

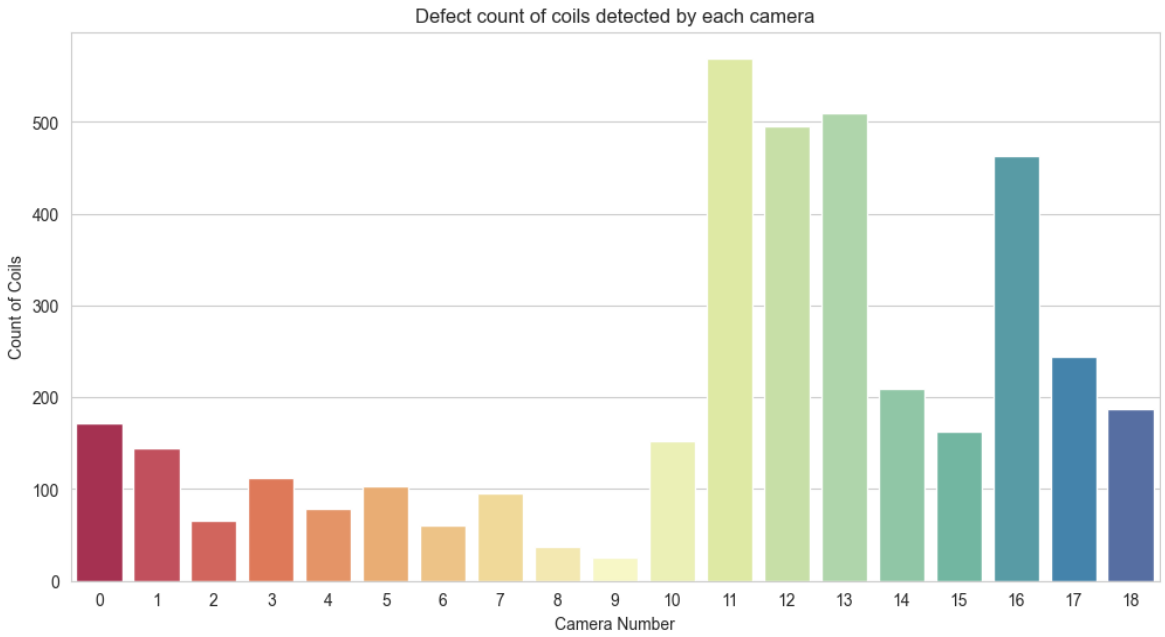
AP4_Ptec_Coils and AP4_Defects_Maps_10_coils and AP4_Defects_Maps_10 _coils merging analysis Report:

Overview

This report encapsulates the data analysis journey undertaken to glean insights from coil and camera-related datasets. By meticulously merging, cleaning, and analyzing data from multiple sources, we have developed a clear understanding of the interaction between camera usage and coil counts in an industrial setting. This analysis is pivotal in optimizing inspection processes and understanding defect detection patterns.

Process Overview

- **Data Integration**
  - Initial Datasets**:** The process began with two key datasets: "AP4_Ptec_Coils" — comprising specific coil data. "AP4_Defects_Maps_10_coils" — detailing defect mapping for a subset of coils.
  - Merging: These datasets were merged, creating a rich, unified source of information that combines coil characteristics with defect mapping by the primary key column("CoilID").
- **Creation of 'AllCoils.csv**
  - Pre merging, the data set underwent further processing. We dropped few columns and converted data columns to effective columns like 'BdeCoilId' to categorical and time columns to datetimeformat.
  - The output was "AllCoils.csv," a dataset offering a more detailed perspective on coil attributes and defects.
- **Deriving 'Cameracoilcounts.csv**
  - Extracting crucial insights from "AllCoils.csv" led to the formation of "Cameracoilcounts.csv."
  - This file captures the relationship between camera usage (for inspections or defect detection) and coil data, summarized through counts.
- **Analysis and Visualization**
  - A visualization is generated from "Cameracoilcounts.csv":
  - Bar Chart: This visualization provided a clear depiction of how coil inspection counts vary across different cameras, highlighting usage patterns and potential workload distribution.



- **Insights and Implications**
    - The analysis revealed patterns in camera usage, suggesting areas for efficiency improvement and potential bias in defect detection.
    - These findings are instrumental in guiding decisions on equipment utilization, maintenance scheduling, and process optimization in coil inspection and defect analysis.
- **Conclusion**

    Through a methodical approach of data merging, cleaning, and analysis, this project has provided valuable insights into the utilization of cameras in coil inspections and defect detections. The visualizations and summaries derived from the "Cameracoilcounts.csv" dataset are pivotal for making informed decisions in industrial processes, emphasizing the importance of data-driven approaches in operational optimization.

Claims_2023-05.csv analysis report:

Objective:

The primary goal is to clean and analyze the dataset "cleaned_claims_2023-05.csv" for better usability and insight extraction. The specific tasks include:

- **Row Filtering**: Exclude rows where the 'LastInspectionMachine' column matches any of the values in ` ['INSP2', 'INSC1', 'INSB2'] `. This step is intended to remove data associated with specific inspection machines, due to known issues or irrelevance to the analysis.
- **Data Cleaning**: We applied general data cleaning techniques, such as stripping whitespace from string columns, to enhance data quality. This step ensures that the text data is consistent and free of leading/trailing spaces that could affect data analysis and processing.
- **Data Inspection and Analysis**: Examine the dataset to understand its structure, identify key columns, and assess data quality. This analysis includes understanding data types, identifying missing or inconsistent data, and getting an overview of the dataset's content.

Process Overview

- **Data Loading**: Import the dataset using pandas, a powerful Python library for data manipulation and analysis.
- **Preliminary Data Inspection**:
    - Assess the structure of the dataset, including the number of rows and columns.
    - Review the first few rows to understand the data format and contents.
    - Identify the data types of each column to determine appropriate processing methods.
- **Data Cleaning and Transformation**:
    - Whitespace Stripping: Strip leading and trailing whitespace from string (object) columns to standardize text data.
- **Output:**
    - Save the cleaned and processed dataset to a new CSV file.

This process aims to make the dataset more manageable and reliable for further analysis or reporting.

tblFlatInspectionMappedDefects and tblFlatInspectionProcess merging analysis report:

Objective: The goal of this analysis was to merge and analyze two key datasets related to flat inspection processes and mapped defects, to understand common defects, quality distributions, and other insights.

Process Overview:

- **Data Loading**:
  - Two datasets were loaded for analysis:
  - **cleaned_tblFlatInspectionMappedDefects.csv**: Contains detailed information about mapped defects.
  - **cleaned_FlatInspectionProcesses.csv**: Includes data on flat inspection processes.
  - These datasets were loaded into Python using the Pandas library.
- **Data Merging**:
  - The two datasets were merged into a single dataset (**merged_df**) based on the **InspectionProcessID** column. This step combined related records from both datasets to form a comprehensive view.
- **Data Cleaning**:
  - Unnecessary columns were removed from the merged dataset to focus on relevant data for analysis.
  - Additional cleaning included stripping whitespace from the **Observations** column.

- **Saving Processed Data**:
  - The cleaned and merged data was saved as a new CSV file, **cleaned_FlatInspectionProcesses.csv**, for further use or reference.

Analysis and Insights Extraction:

- **Time Analysis**: The code analyzed the inspection processes, particularly focusing on the time taken for inspections. It compared inspection times for processes with and without defects, converting these times into a human-readable format (hours, minutes, seconds).

- **Defect Analysis**:
  - The most common defect types were identified and counted.
  - The relationship between defects and the quality of processes was explored by grouping and counting defects based on quality descriptions.
  - A similar analysis was done to understand the relationship between defects and Line IDs.
- These analyses were saved as separate CSV files for easy access and reference.

```
Quality Correlation - Defects by Quality:

Quality: First
Description
STREAKY ROUGHNESS                                  290
Rolled in Marks / Debris from scalebreaker AP4     125
BRIDLE ROLL MARK AP4-255"/1-152"/2-225"/3-99"      121
OIL STAINS                                         111
STAINS FROM ALKALI SECTION                          98
Name: count, dtype: int64

Defect and Process Relationship - Defects by LineID:

LineID: 25
Description
WORK ROLL STOP FROM Z.M.                                  55
OP/INSP OUT PERFORMING OPERATION DUTIES                   47
FURNACE ROLL MARK FROM HOT AND/OR COLD AP (27-32")        38
Veiny Acid Stains                                         37
OIL STAINS                                                37
Name: count, dtype: int64
```

Visualization:

- Two key visualizations were generated:
- A bar chart showing the top 10 most common defect types, providing a clear picture of the most frequent issues encountered.
- A bar chart displaying the count of different quality descriptions in the dataset, giving an overview of quality distribution.

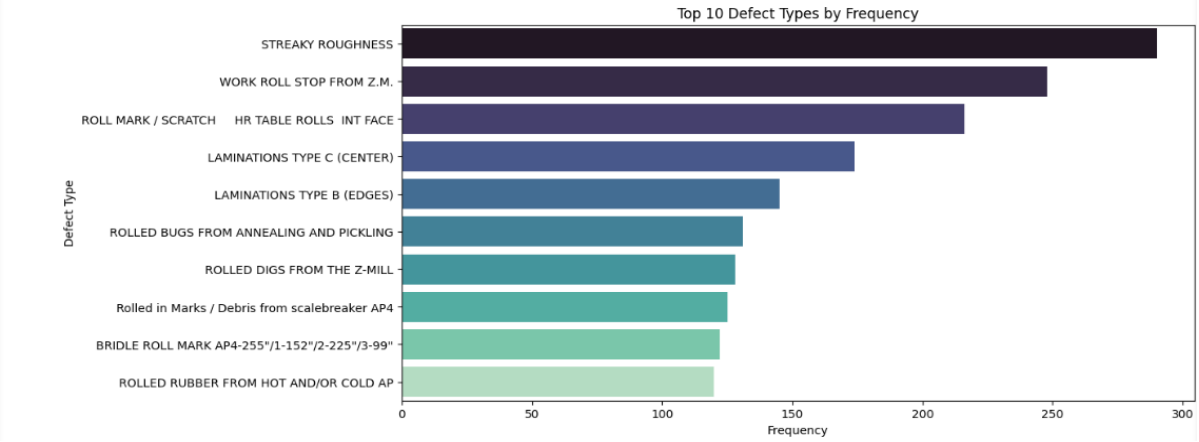


Fig: Top ten Defect types by frequency

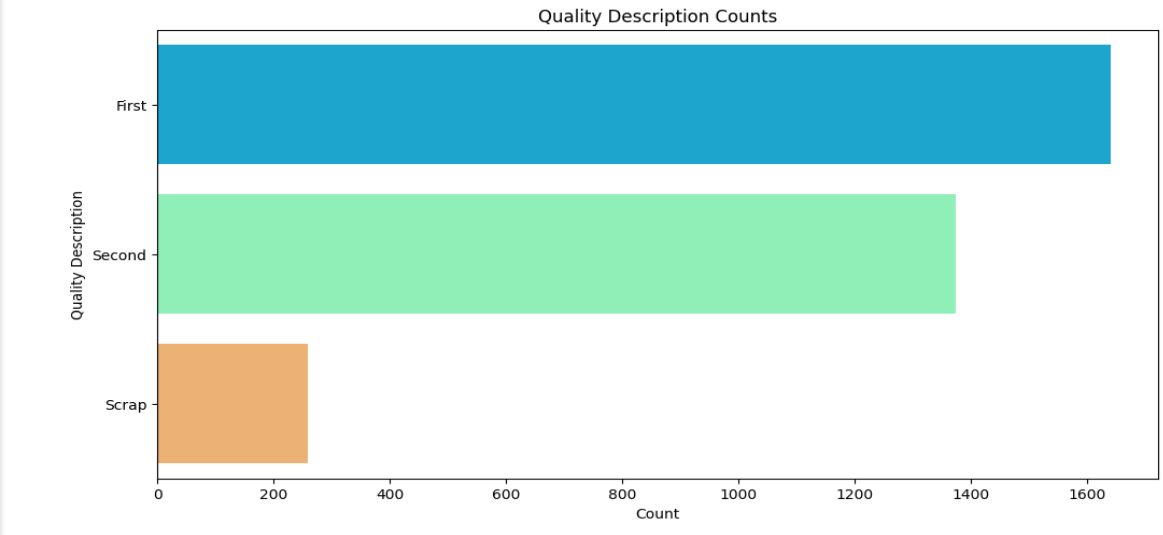


Fig: Quality Description counts

Conclusion: This analysis provided a detailed view of the inspection processes and defects, highlighting common defect types and quality trends. The visualizations offered an easy-to-understand representation of these key aspects, aiding in quick comprehension and decision-making.

FLInspectionComments and tblFLInspection merging analysis Report:

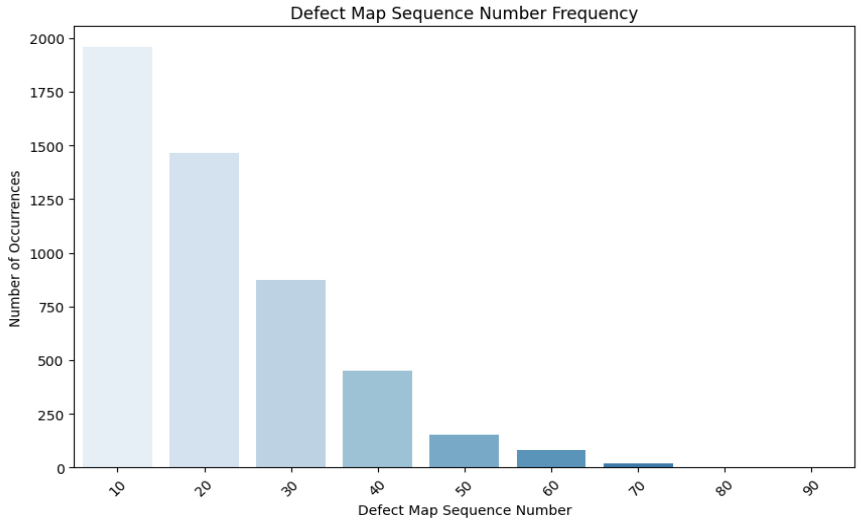
Objective

The code aims to merge two datasets related to facility inspections, and then analyze the frequency of different defect map sequence numbers within the merged data.

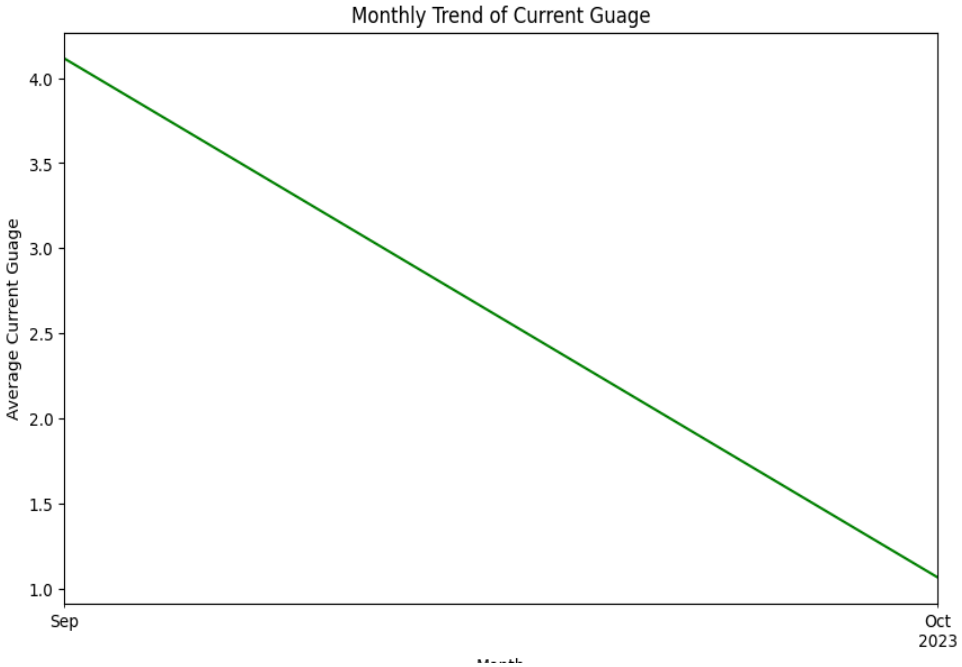
Process Overview

- **Data Preparation**
  - **Datasets Loaded**: Two CSV files, `cleaned_FLInspectionComments.csv` and `cleaned_FlInspection.csv`, are loaded into pandas Data Frames named `comments` and `newdf`, respectively.
- **Merging Process:**
  - Unique `FLInspectionID` values are extracted from the `comments` DataFrame.
  - A filter is applied to `newdf` to keep only those rows where `FLInspectionID` matches with IDs in `comments`.
  - The `comments` and filtered `newdf` DataFrames are merged on `FLInspectionID`. The inner joint ensures only matching records are kept.
- **Output:** A merged Data Frame `merged_df` is created and saved as `InspectionsMerged.csv`.
- **Visualization**

- Provides insights into common defects or areas that may require further investigation or improvement.



- The plot shows how the average 'CurrentGuage' changes over time on a monthly basis. This can reveal underlying trends such as increasing or decreasing patterns, periodic fluctuations, or stability over time.



Observations and Suggestions

- **Data Merging**: The merging strategy is robust, ensuring that only relevant records from both datasets are included in the analysis.
- **Visualization:**
    - The bar plot provides a clear visual representation of the data, aiding in the identification of common or rare defective sequences.
    - Dynamic Visualization: If the dataset is large, we can consider plotting only the top N categories or using a log scale for better visualization.
- **Potential Enhancements:**
    - Error Handling: Incorporating error handling for file reading and merging processes could make the code more robust, especially if there is a chance of encountering corrupt or missing data.
    - Data Exploration**:** Before visualizing, exploring the data with descriptive statistics or checking for null values could provide deeper insights and ensure data quality.

Conclusion

The code effectively merges and analyzes specific aspects of inspection-related data, providing valuable insights into the distribution of defect types or occurrences. This can aid in informed decision-making and strategic planning in quality control and regulatory compliance scenarios.

Dataframes to Database:

In our project, we leveraged the Psycopg2 library for efficient database interactions. The first step involves establishing a connection to the PostgreSQL database using the provided connection parameters.

Next, we dynamically determine the data types for each column in the DataFrame, classifying them as 'BOOLEAN' for boolean types and 'TEXT' for other data types.With the necessary information gathered, we proceed to create a table in the PostgreSQL database using the specified table name and column definitions.

Once the table structure is defined, data is inserted into the table using an appropriate SQL query. Subsequently, changes are committed to the database.Finally, to ensure proper resource management, the cursor and database connection are closed.This systematic approach ensures the seamless integration of data from a DataFrame into a PostgreSQL database, following best practices for connection management and data type handling.

The tables created were: All Coils Table, Merged Coils Table, Camera Count Table, Inspections Merged Table, Processes and Defects Merged Table and claims table.

Issues to address in this Data Analysis Project

1. **Standardization and Consistency in Data Formats:** Across different datasets, inconsistencies in formats (e.g., date-time formats, text casing) can lead to challenges in data merging and analysis. Implementing a standard format for all datasets is crucial.

2. **Handling of Missing or Incomplete Data**: Each analysis report encountered missing or incomplete data. Developing a consistent strategy for handling such data (e.g., imputation, exclusion) is essential to maintain the integrity of the analysis.

3. **Robust Data Cleaning Mechanisms:** While basic cleaning like whitespace stripping is mentioned, more robust cleaning processes (e.g., outlier detection handling of duplicate entries) must be established to ensure data quality.

4. **Enhanced Error Handling and Data Validation:** Particularly in the merging processes, there is a need for better error handling to manage issues like mismatched data types, corrupted files, or inconsistent primary keys across datasets.

5. **Scalability of Data Processing Techniques**: As datasets grow, the current data processing and analysis techniques might need to scale more efficiently. Implementing more scalable solutions, such as using more efficient data processing libraries or parallel processing, is necessary.

6. **Advanced Analytical Techniques and Machine Learning Integration:** The reports primarily focus on descriptive analytics. Incorporating advanced analytical techniques, such as predictive modeling and machine learning, could provide deeper insights and more actionable outcomes.

7. **Dynamic and Interactive Visualization Tools:** While static visualizations provide initial insights, there is a scope for implementing more dynamic and interactive visualizations like those generated using PowerBi. These tools can offer a more in-depth and customizable analysis experience, especially for large and complex datasets.

Addressing these issues will significantly enhance the data analysis, leading to more accurate, reliable, and insightful outcomes.

Lesson learned

1. **Importance of Comprehensive Data Cleaning**: The analysis reinforced the critical role of thorough data cleaning in ensuring accurate results. Lessons include understanding the nuances of the data, the significance of removing outliers, standardizing formats, and dealing with missing values to maintain the integrity of the analysis.

2. **Effective Data Integration Techniques**: Merging datasets from various sources highlighted the need for effective data integration strategies. This experience teaches the value of identifying and using appropriate key columns for merging, addressing data inconsistencies, and ensuring data compatibility across various sources.

3. **Scalability and Efficiency in Data Processing**: Dealing with large datasets underscored the importance of scalability and efficiency. Lessons learned involve using more efficient data processing methods, such as optimized algorithms or parallel processing, to manage growing amounts of data while maintaining optimal performance.

4. **Advanced Analytical and Predictive Modeling Skills**: The projects demonstrated the limitations of purely descriptive analytics and the potential of advanced analytical techniques. Incorporating predictive modeling and machine learning provides deeper insights and aids in forecasting and strategic decision-making.

5. **Value of Dynamic and Interactive Visualizations**: Static visualizations have their limits. The analysis taught the importance of dynamic and interactive visualization tools for more engaging and insightful presentations, especially when dealing with complex datasets.

6. **Robust Error Handling and Data Validation**: Encountering various data issues highlighted the necessity for robust error handling and validation mechanisms. These lessons include developing strategies to anticipate and manage common data issues, such as format inconsistencies, corrupt files, and unexpected data values, to ensure the reliability of the analysis process.

Future data analysis projects can achieve higher accuracy, efficiency, and impact by incorporating these lessons.