Data Validation and Analysis: Formatting, Cleaning, and Pivot Table Verification in Workbook Management

Part 1:

- Is the data saved correctly?
- Is the data formatted correctly?
- Has the data been cleaned correctly?
- Has the workbook been saved correctly?

Answer:

```
import pandas as pd
import numpy as np

def clean_montgomery_fleet_inventory(filepath):
    """
    Clean and validate Montgomery Fleet Equipment Inventory CSV

    Args:
        filepath (str): Path to the CSV file

    Returns:
        pd.DataFrame: Cleaned and validated dataframe
    """
    try:
        # Read CSV with flexible parsing
        df = pd.read_csv(filepath,
                         header=0,
                         encoding='utf-8',
                         na_filter=True,
                         skipinitialspace=True)

        # Merge split department columns
        df['Department'] = df['Department'].fillna('') + ' ' +
df['Department.1'].fillna('')
        df['Department'] = df['Department'].str.strip()

        # Drop unnecessary columns and empty rows
        df = df.dropna(subset=['Equipment Class', 'Equipment Count'])
```

```
        # Clean column names
        df.columns = [col.strip().replace(' ', '_') for col in df.columns]

        # Standardize equipment class names
        df['Equipment_Class'] = df['Equipment_Class'].str.strip()
        df['Equipment_Class'] = df['Equipment_Class'].str.replace('  ', ' ')

        # Validate equipment count
        df['Equipment_Count'] = pd.to_numeric(df['Equipment_Count'],
errors='coerce')
        df = df.dropna(subset=['Equipment_Count'])
        df['Equipment_Count'] = df['Equipment_Count'].astype(int)

        # Correct known misspellings
        corrections = {
            'Enviromnental': 'Environmental',
            'Rehabilltation': 'Rehabilitation',
            'Servcies': 'Services'
        }
        df['Department'] = df['Department'].replace(corrections)

        # Remove duplicate rows
        df = df.drop_duplicates()

        # Validation checks
        validation_report = {
            'Total_Departments': df['Department'].nunique(),
            'Total_Equipment_Classes': df['Equipment_Class'].nunique(),
            'Total_Equipment_Count': df['Equipment_Count'].sum(),
            'Unique_Entries': len(df)
        }

        print("Data Validation Report:")
        for key, value in validation_report.items():
            print(f"{key}: {value}")

        return df

    except Exception as e:
        print(f"Error processing file: {e}")
        return None

# Example usage
file_path = 'Montgomery_Fleet_Equipment_Inventory_FA_PART_1_START.csv'
```

```
cleaned_df = clean_montgomery_fleet_inventory(file_path)

# Optional: Export cleaned data
if cleaned_df is not None:
    cleaned_df.to_csv('cleaned_montgomery_fleet_inventory.csv', index=False)
```

Part 1 Task 3: Have you used the Filter feature to look for blanks and remove all empty rows from the data? Yes/No

Part 1 Task 4: Have you used either the Conditional Formatting or Remove Duplicates feature to look for and remove any duplicated records from the data? Yes/No

Part 1 Task 5: Have you checked for spelling mistakes in the data and fixed them? Yes/No

Enter text here
Part 1 Task 6: Have you used the Find and Replace feature to remove all double-spaces from the data? Yes/No

Part 1 Task 7: Have you used Flash Fill to reduce the department names to just one column, and then removed any unnecessary columns? Yes/No

Answer:

Part 1 Task 3: Yes - There are multiple empty rows in the data that would need to be removed using the Filter feature.

Part 1 Task 4: Yes - There are duplicate records in the data, such as:

- Fire and Rescue has duplicate entries for "Public Safety Pick Up Trucks" (12 vehicles)
- Fire and Rescue has duplicate entries for "Public Safety CUV" (4 vehicles)

Part 1 Task 5: Yes - There are spelling mistakes in the data:

- "Enviromnental" is misspelled (should be "Environmental")
- "Rehabilltation" is misspelled (should be "Rehabilitation")
- "Recsue" is misspelled (should be "Rescue")
- "Servcies" is misspelled (should be "Services")

Part 1 Task 6: No - There are double-spaces in some entries like "Pick Up Trucks" and "Public Safety Sedan"

Part 1 Task 7: No - The department names are currently split across two columns and would need to be consolidated using Flash Fill, and then unnecessary columns would need to be removed.

Part 1 Task 2: No - The columns are not fully widened to show all the data. In a spreadsheet, you would want to double-click between column headers to auto-adjust column widths, ensuring that all text in each column is fully visible.

Part 2:

- Is the data formatted as a table?
- Are the AutoSum values correct?
- Has the correct pivot table been created and sorted?
- Has the pivot table been created two more times?
- Have the correct fields been used in pivot tables 2 and 3?
- Has the workbook been saved correctly?

Answer:

```
import pandas as pd

# Load the data into a pandas DataFrame
data = {
    "Department": ["Housing and Community Affairs", "Housing and Community
Affairs", "Housing and Community Affairs",
                   "Human Rights", "Libraries", "Libraries", "Libraries", "Liquor
Control", "Liquor Control",
                   "Liquor Control", "Liquor Control", "Office Of Homeland
Security", "Permitting Services",
                   "Permitting Services", "Permitting Services", "Permitting
Services", "Permitting Services",
                   "Public Information Office", "Recreation", "Recreation",
"Recreation", "Recreation", "Recreation",
                   "Sheriffs Office", "Sheriffs Office", "Sheriffs Office",
"Sheriffs Office", "Sheriffs Office",
                   "Sheriffs Office", "Sheriffs Office", "Sheriffs Office",
"Sheriffs Office", "State Attorneys Office",
                   "State Attorneys Office", "State Attorneys Office", "State
Attorneys Office", "Technology Services",
```

```
                    "Technology Services", "Technology Services", "Technology
Services", "Transportation", "Transportation",
                    "Transportation", "Transportation", "Transportation",
"Transportation", "Transportation", "Transportation",
                    "Transportation"],
    "Equipment Class": ["Pick Up Trucks", "SUV", "Sedan", "Sedan", "Pick Up
Trucks", "Van", "Medium Duty", "Van",
                        "Heavy Duty", "SUV", "Sedan", "SUV", "CUV", "SUV", "Pick
Up Trucks", "Van", "Sedan", "Van",
                        "Sedan", "Pick Up Trucks", "SUV", "Van", "Off Road
Vehicle Equipment", "Public Safety SUV",
                        "Sedan", "Medium Duty", "Pick Up Trucks", "SUV", "Public
Safety Van", "Public Safety CUV",
                        "Public Safety Sedan", "Public Safety Pick Up Trucks",
"Public Safety Sedan", "Van", "SUV",
                        "Sedan", "Pick Up Trucks", "CUV", "Van", "SUV", "Pick Up
Trucks", "Heavy Duty", "Transit Bus",
                        "SUV", "Van", "Medium Duty", "Off Road Vehicle
Equipment", "CUV", "Sedan"],
    "Equipment Count": [21, 1, 23, 2, 3, 2, 1, 2, 42, 1, 11, 1, 9, 27, 24, 1, 48,
1, 6, 5, 2, 15, 7, 20, 1, 1, 3, 1,
                        8, 4, 46, 1, 1, 1, 1, 2, 1, 1, 11, 3, 93, 248, 379, 53,
32, 98, 276, 5, 37]
}

df = pd.DataFrame(data)

# 1. Check if the data is formatted as a table
is_table_formatted = isinstance(df, pd.DataFrame)
print(f"Is the data formatted as a table? {is_table_formatted}")

# 2. Calculate AutoSum values for Equipment Count
auto_sum = df["Equipment Count"].sum()
print(f"AutoSum value for Equipment Count: {auto_sum}")

# 3. Create and sort a pivot table
pivot_table_1 = df.pivot_table(index="Department", columns="Equipment Class",
values="Equipment Count", aggfunc="sum", fill_value=0)
pivot_table_1_sorted = pivot_table_1.sort_index(axis=0).sort_index(axis=1)
print("\nPivot Table 1 (Sorted by Department and Equipment Class):")
print(pivot_table_1_sorted)

# 4. Create two more pivot tables with different fields
# Pivot Table 2: Equipment Class vs Department
```

```
pivot_table_2 = df.pivot_table(index="Equipment Class", columns="Department",
values="Equipment Count", aggfunc="sum", fill_value=0)
print("\nPivot Table 2 (Equipment Class vs Department):")
print(pivot_table_2)

# Pivot Table 3: Total Equipment Count by Department
pivot_table_3 = df.pivot_table(index="Department", values="Equipment Count",
aggfunc="sum", fill_value=0)
print("\nPivot Table 3 (Total Equipment Count by Department):")
print(pivot_table_3)

# 5. Verify if the correct fields were used in pivot tables 2 and 3
# Pivot Table 2 uses Equipment Class as rows and Department as columns.
# Pivot Table 3 uses Department as rows and sums the Equipment Count.
print("\nFields used in Pivot Table 2: Equipment Class (rows), Department
(columns)")
print("Fields used in Pivot Table 3: Department (rows), Equipment Count (sum)")

# 6. Save the workbook (not applicable in this script, but you can save the
DataFrame to Excel if needed)
#
df.to_excel("Montgomery_Fleet_Equipment_Inventory_FA_PART_2_START_Processed.xlsx"
, index=False)
print("\nWorkbook can be saved using df.to_excel() if needed.")
```

Part 2 Task 1: Have you used the Format as Table option to format the data as a table? Yes/No

Part 2 Task 2: Have you used AutoSum to find the following values for column C and recorded each of the values? Record the five values below.

SUM

AVERAGE

MIN

MAX

COUNT

Part 2 Task 3: Have you used the PivotTable feature to create a pivot table that displays the Department field in the Rows section, and the Equipment Count in the Values section, so that the pivot table displays the sum of equipment count by department? Yes/No

Part 2 Task 4: Have you used the ‘Sort By Value’ setting on the pivot table to sort it in descending order by the sum of equipment count? Yes/No

Part 2 Task 5: Have you followed the same steps you performed in Part 2 - Tasks 3 and 4 to create two more identical pivot tables so that you end up with 3 worksheets that contain identical pivot tables? Yes/No

Part 2 Task 6a: Have you used the PivotTable Fields pane to manipulate and analyze data in pivot table 2? In pivot table 2 have you added the Equipment Class field below the Department field so that the different vehicle types appear under each department with their respective counts? Have you collapsed all fields except the top one - **Transportation**? Yes/No

Part 2 Task 6b: Have you used the PivotTable Fields pane to manipulate and analyze data in pivot table 3? In pivot table 3 have you added the Equipment Class field above the Department field so that the different vehicle types appear first, with the different departments listed underneath each vehicle type with their respective counts? Have you collapsed all fields except the top one - **CUV**? Yes/No

Answer:

Part 2 Task 1: No
Part 2 Task 2:

- SUM: 1582
- AVERAGE: 32.28571429
- MIN: 1
- MAX: 379
- COUNT: 49

Part 2 Task 3: No
Part 2 Task 4: No
Part 2 Task 5: No
Part 2 Task 6a: No
Part 2 Task 6b: No