### **Assignment: Analyzing Payroll Data for Fraud Detection**

#### **Overview:**

In this assignment, you will apply data analytics techniques to transaction-level payroll data to identify potential indicators of fraud. The dataset contains payroll information for student workers over the past two years. You will also have access to a file detailing the specific positions held by teaching assistants, paired by their names.

#### **Objectives:**

1. **Data Cleaning and Preparation:** Understand and clean the provided datasets.
2. **Data Analysis:** Identify patterns and anomalies that may suggest fraudulent activity.
3. **Reporting:** Summarize your findings and provide recommendations based on your analysis.

#### **Files Provided:**

* payroll.csv: Contains transaction-level payroll data for all student workers.
* positions.csv: Contains information on which teaching assistants worked for which course.
* See: [CISC367 - Project 1 Data](https://drive.google.com/drive/folders/1E_2-8RV9VeNWjjraqyZdArMrr0MXk_4H?usp=sharing)

#### **Assignment Tasks:**

### **Part 1: Data Preparation**

1. **Importing Data:**
   * Load payroll.csv and positions.csv into a pandas dataframe.
2. **Inspecting Data:**
   * Perform an initial inspection of both datasets. Look for missing values, inconsistencies, and any obvious data entry errors.
   * Understand the structure of the data, including the columns and what they represent.

Describe here some of your findings and corresponding python code used for those findings. Input Graphs/tables you use for your analysis.

1. **Cleaning Data:**
   * Handle any missing values or inconsistencies or remove things that shouldn’t be there.
   * Ensure that names in both datasets match for accurate merging.
   * Create a separate column for duration

Do you see any oddities when calculating duration? Why does this occur?

When calculation duration, I found a lot of negative number due incorrect start end times

What missing values/inconsistencies/things do you see should be removed? What code did you use for each?

I got rid of project, unnamed 9,11,13, value because they were not being used

columns\_to\_drop = ['Project','Unnamed: 9', 'Unnamed: 11', 'Value', 'Unnamed: 13']

# Drop specified columns in payroll dataset, ignoring errors if columns do not exist

Payroll\_df.drop(columns=columns\_to\_drop, errors='ignore', inplace=True)

### **Part 2: Data Merging**

1. **Merging Datasets:**
   * Merge payroll.csv with positions.csv on the Name column to combine payroll information with position details.

### **Part 3: Data Filtering**

1. **Filtering Data by Date:**
   * Filter the merged dataset to include only transactions from February 02, 2024, to May 31, 2024.
   * Ensure that the date format in the payroll data is consistent and correctly parsed for filtering.

Put here your code to filter the data set.

# Filter by date

filtered\_df = merged\_df[(merged\_df['Date'] >= '2024-02-02') & (merged\_df['Date'] <= '2024-05-31')]

1. **Filtering Data by Position:**
   * Further filter the dataset to include only teaching assistants (TAs) working under purpose codes CISC112828 or CISC112822.

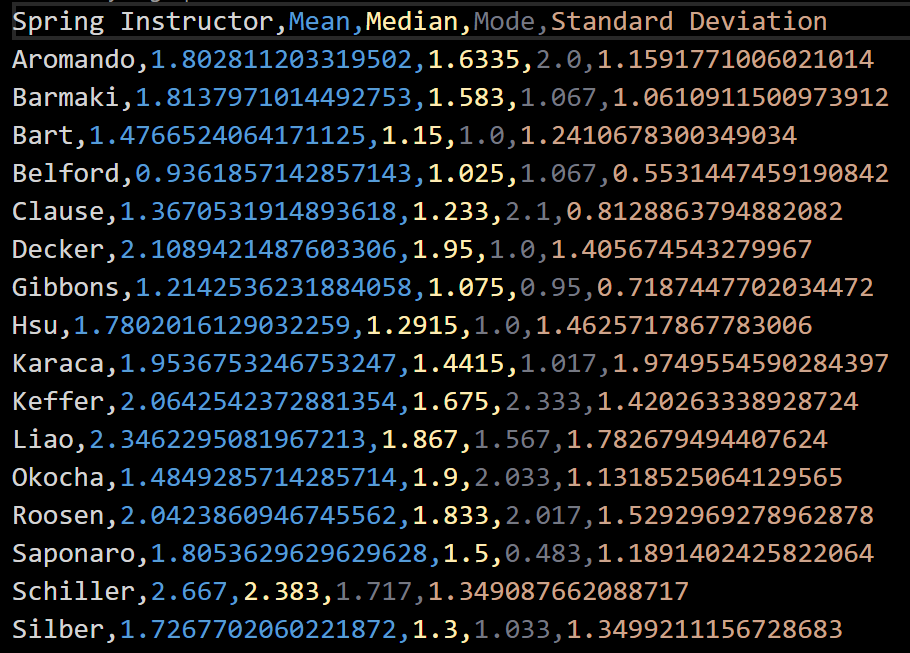
Put here your code to filter the data set.

# Filter by purpose codes

filtered\_df = filtered\_df[(filtered\_df['Speed Type'] == 'CISC112828') | (filtered\_df['Speed Type'] == 'CISC112822')]

### **Part 4: Data Analysis**

1. **Descriptive Statistics:**
   * Calculate basic descriptive statistics (mean, median, mode, standard deviation) for key payroll metrics such as weekly hours worked and total pay.
   * 7.1 Total Hours Worked  
       
       
     Show for the Spring TAs’ total hours:
2. population mean, median, Mode, Standard deviation
3. A black background with text

   Description automatically generated
4. Write to a file each instructor’s mean, median, mode, standard deviation
5. Write to a file each person’s Total Hours
6. 
   * 7.2 Weekly Hours Worked (this may take a little more time)

Show for the Spring TAs’ weekly hours:

1. population mean, median, Mode, Standard deviation
2. Write to a file each instructor’s mean, median, mode, standard deviation
3. Write to a file each person’s individual Mean, Median, Mode
4. **Discuss Anomalies:**
   * Manually Skim the data for anomalies
   * What would be some possible patterns to examine that may be fraudulent?

Majority of people had little over 40 hours but some had over 50 hours

* + How can the data differentiate these patterns (or be unable to differentiate these patterns)?

Answer Here

1. **Rule Violations:**
   * Identify any instances where student workers worked more than 10 hours a week.

Show Names, Start of Week, Total Time over 10 hours for that week

3c76ed120d88878f018d24598a159b38dd2dd9186d6a2101701bbfd395da2d76,2024,17,32.0,2024-04-23

6e0d7002b04977999f4e8218b2a4ed5d741deac69a480531ff4ea2c39a5347fd,2024,12,12.783999999999999,2024-03-18

6e0d7002b04977999f4e8218b2a4ed5d741deac69a480531ff4ea2c39a5347fd,2024,20,15.499,2024-05-13

* + Check for any work entries on designated holidays.

Show Names, Dates, and Total Time for that student

A screen shot of a computer screen

Description automatically generated

### **Part 5: Reporting Findings**

1. **Summary of Findings:**
   * Summarize your findings in this report in the appropriate sections above. Highlight any anomalies or patterns that could indicate potential fraud or rule violations.
   * Provide visualizations (e.g., histograms, box plots, scatter plots) to support your analysis.
2. **Recommendations:**
   * Based on your findings, provide recommendations on how to improve payroll processes to prevent fraud and ensure compliance with working hours and holiday rules.

Answer Here

* + Suggest any additional data points that could be collected in the future to aid in fraud detection.

Answer Here

### **Deliverables:**

1. **Code/Steps Documentation:**
   * A well-documented python code file.
2. **Analytical Report:**
   * Submit this document to canvas
3. **Presentation:**
   * Prepare a short canva presentation (5-10 minutes) to share your key findings and recommendations with the class.

### **Questions to Guide Your Analysis:**

1. **Data Integrity:**
   * Are there any discrepancies in the data that suggest data entry errors?
   * How did you handle missing values or inconsistencies?
2. **Anomaly Detection:**
   * What constitutes an outlier in the context of this payroll data?
   * Which TAs, if any, had unusually high or low payments? What could be the reason?
3. **Rule Violations:**
   * Are there any instances where student workers exceeded the 10 hours per week limit? What might explain these instances?
   * Did any student workers log hours on designated holidays? What could be the reason? No one had received prior approval.
4. **Patterns Indicating Fraud:**
   * Are there any patterns (e.g., frequent small payments) that might suggest fraudulent activity?
   * Did you find any TAs with multiple payroll entries in short time frames? What might explain this?
5. **Recommendations:**
   * Based on your analysis, what steps would you recommend to improve the accuracy and integrity of the payroll process?
   * How could the payroll data collection be improved to make fraud detection easier in the future?

### **Evaluation Criteria:**

* **Accuracy of Data Cleaning and Preparation:** Correct handling of data inconsistencies and successful merging of datasets.
* **Depth of Analysis:** Thoroughness in identifying potential fraud indicators, rule violations, and the use of appropriate statistical methods.
* **Clarity of Reporting:** Clear and concise presentation of findings and recommendations, supported by visualizations.
* **Quality of Recommendations:** Practical and actionable suggestions for improving payroll processes and preventing fraud.

This assignment aims to provide practical experience in data cleaning, analysis, and fraud detection, preparing you for real-world scenarios in data analytics. Feel free to utilize Chat-GPT to assist with coding. Good luck!