CGT 270 Data Visualization

Module 1

Week 3

**Lab 3: Mining Data**

The goal of this lab is to identify and implement techniques for mining data. In this lab you will identify patterns, extreme and subtle feature about data. You will identify basic descriptors for the data, and categorize data according to the specifications defined in the Parse Worksheet you completed in Week 2. After completing this lab, you will:

1. List at least three (3) questions you feel you can answer with the data sets you have acquired (Week 1) and parsed (Week 2).
2. Your questions must incorporate ALL three (3) of the data sets you’ve acquired from Lab 1: Tableau Dataset, Additional Dataset #1, and Additional Dataset #2
3. List any assumptions you are making in this stage of the data visualization process.

**What you should be able to do (at the end of this lab):**

|  |  |
| --- | --- |
| Understand | ***Describe*** the type of techniques to be used to better understand the data. |
| Apply | ***Execute*** techniques and methods (statistical methods) on the data. |
| Evaluate | ***Examine*** the resulting data and determine if it enables you to answer the question being solved. |
| Analysis | ***Identify*** patterns, extreme and subtle features about the data. |
| Create | ***Determine*** if the data can support the question to be answered. |

In the table below list each variable in the Tableau dataset, its data type (parsing) and a basic statistical or mining technique that can be applied to better understand the variable.

**Part I: Tableau Data set:** *Global Sport Finances: Top Athlete Salaries*

1. **Basic Descriptors**

List the **variables** from Week 2’s parsing lab and provide basic mining procedures.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Type** | **Basic mining procedure** |
| Name | String | String length |
| Pay | Integer | Max, Min, Average |
| Salary/Winnings | Integer | Max, Min, Average |
| Endorsements | Integer | Max, Min, Average |
| Sport | String | String Length |

Add more rows to the table above as needed.

1. **Categorize**

Consider what variables are similar and what variables are different. This will help you to categorize the data. Are the data nominal, ordinal or ratio? Take a look at this webpage and video: <https://www.graphpad.com/support/faq/what-is-the-difference-between-ordinal-interval-and-ratio-variables-why-should-i-care/>

Nominal: Sport type, name of athletes

Ordinal: Pay, Salary/winnings, Endorsements

Ratio: Salary/Winnings + Endorsements : Pay

Review the different types of data and indicate the data types in your variables table:

<https://www.centralriversaea.org/wp-content/uploads/2017/03/F_Four-Types-of-Data-Revised-5.10.17.pdf>

1. **Temporal**

Is the data temporal (represent time, over several years, in years, days, minutes, seconds)?

Yes, this data set represents just the time between January, 2014 to December, 2014.

1. **Range and Distribution**

What is the distribution of the data? Few values, small size, evenly spread, sparse or dense? Explain.

The dataset is small with only 101 rows. The data seems to be evenly spread with a range of 20 million to 50 million, with an outlier of 105 million. The average of the data is 61.5 million.

**Part II: First (1st) additional data set:** *Forbes Richest Athletes 1990-2020*

1. **Basic Descriptors**

List the variables from Week 2’s parsing lab and provide basic mining procedures.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Type** | **Basic mining procedure** |
| Name | String | String length |
| Nationality | String | String length |
| Rank | Integer | Max, Min |
| Previous year ranking | Integer | Max, Min |
| Sport | String | String length |
| Year | Integer | Max, Min |
| Earnings | Integer | Max, Min, Average |
|  |  |  |

Add more rows to the table above as needed.

**Part III: Second (2nd) additional data set:** *NBA Player Salaries 2000-2019*

1. **Basic Descriptors**

List the variables from Week 2’s parsing lab and provide basic mining procedures.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Type** | **Basic mining procedure** |
| Name | String | String length |
| Year | Integer | Max, Min |
| Salary | Integer | Max, Min, Average |
| Rank | Integer | Max, Min |
|  |  |  |

Add more rows to the table above as needed.

**Part IV: Questions and Assumptions**

List at least three (3) questions you feel you can answer using the datasets you have acquired and mined. You MUST use complete sentences. Your questions must incorporate ALL three (3) of the data sets you’ve acquired.

Q1: What was the Average Salary of an NBA player from the year 2000-2019?

Q2: Which athlete made the highest amount in endorsements in the year 2014?

Q3: What sport did the richest athlete played from the year 1990-2020?

**List 3 assumptions you are making in this stage of the data visualization process:**

1. **Assumption #1:** Salaries listed in this data set are accurate and are not rounded.
2. **Assumption #2:** The dataset took into consideration both male and female athletes in their respective sports.
3. **Assumption #3:** The dataset took all of the athletes sources of income into consideration, when determining the yearly salary of the athlete.