**Project 3.2.6 - Investigating with Data**

With a partner, select some questions you would like to investigate. Use relevant data and data analysis tools to create meaningful insight into your area of investigation. Then create a visualization that summarizes the information/knowledge you gained from your analysis of the data.

**Part 1**

Develop a cluster of questions in an area of investigation. To develop these questions, use the brainstorming activities below and record your answers in the Project 3.2.6 - Brainstorming Log.

1. Brainstorm areas in which you might be interested in creating new insight or knowledge. You can include broad and more specific areas in your brainstorming. For example, you might list music, live music, and/or a genre of music as areas for investigation. Recall the guidelines for brainstorming: use a tagline to keep a record, go for quantity, never criticize ideas during brainstorming, and “piling on” is welcome.
2. Brainstorm fields of data that might be obtainable for each of two or more of the areas from the previous brainstorm. For example, for orchestral music, there might be data listing composer, piece, year, conductor, orchestra, date of performance, location, and so on. Keep a record of this second brainstorm.
3. Brainstorm questions that ask how one field of data is related to another field of data. For example, how does the location in which an orchestral music piece was composed relate to the location in which it has been performed? Has the “export” of music changed over time?
4. Select 3-5 questions in a single area of investigation. Describe the area of investigation and why you chose it.

**Part 2**

Seek one or more sets of data relevant to your investigation. The combined data set should contain multiple attributes and be rich enough to provide meaningful insight in your area of investigation. You can proceed with the next step while continuing to search for additional data.

1. Identify each data set and its source. Use an APA citation with a URL and the date of retrieval.
2. Describe each data set.

**Part 3**

Use any tools you wish to seek patterns in your data. Consider working with subsets of the data by sorting, clustering, and filtering the data. Comment any code you produce. Use visualization and analysis as appropriate, recording and explaining your work as you progress. Keep records of your methods, results, and discoveries in a project notebook. Keep version-controlled copies of your code and derived data.

**Part 4**

After you and your partner have finished investigating your data, individually complete parts A through E below.

1. What questions did you select to investigate and what was your rationale for selecting these questions.

The questions I decided to investigate was ‘Does the most watched sport receive the most income?’, ‘Does the amount of people who play the sport correlate to how many watch it?’, and lastly ‘Does the least amount of injuries relate to how many play he sport?’ When choosing the field of data for the topic of sports, I thought that how many injuries, most watched sport, and the income in a specific sport were the most obtainable pieces of data I could find online. After careful deliberation and connecting the possible fields of data the questions were created. The question that was most interesting was ‘Does the most watched sport receive the most income?’ This question began our search for data because money, through it can not buy happiness, is an important part of our everyday lives and sports are the first kind of entertainment that appeal to a wide age range of individuals. Together, the question for the data was formed.

1. Provide APA citations and descriptions of the data you used.

Sports, O. S. (2016, November 13). Baseball Databank. Retrieved February 5, 2019, from <https://www.kaggle.com/open-source-sports/baseball-databank#Salaries.csv>

The data from kaggle.com for baseball provided data on the baseball players, teams, games, and salaries from 1871- 2015 as a table.

Crawford, C. (2017, July 13). U.S. Major League Soccer Salaries. Retrieved February 5, 2019, from [https://www.kaggle.com/crawford/us-major-league-soccer-salaries#mls-salaries-2015.csv](https://www.kaggle.com/crawford/us-major-league-soccer-salaries#mls-salaries-2017.csv)

The data from kaggle.com for soccer provided data on soccer players from 2015 in the United States, their clubs, their positions, and their salaries as a table.

Shoup, S. (n.d.). 10 Most Popular Sports Americans Love. Retrieved February 5, 2019, from <https://fanspeak.com/steveospeak/2017/04/01/10-popular-sports-americans-love/>

The data on fanspeak.com provided a list of the top 10 sports Americans watch the most. Their data was created by conducting a survey.

1. Explain the information or knowledge you gained from your analysis of the data.

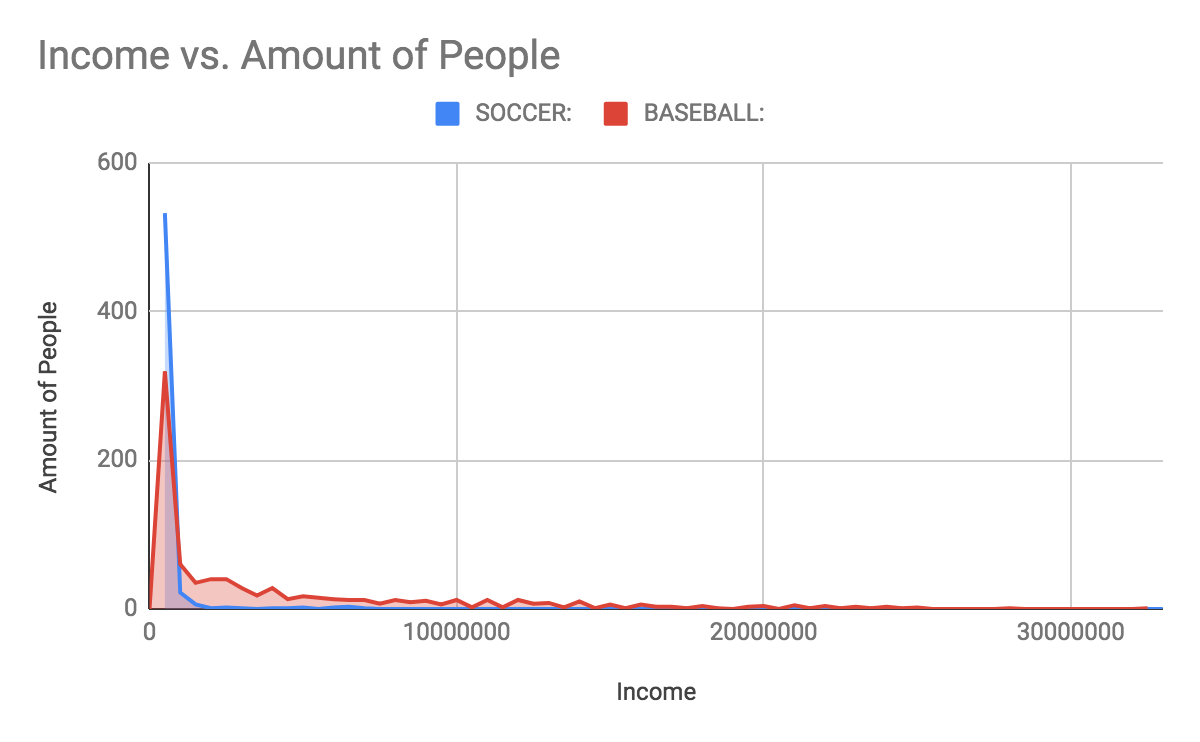
From the fanspeak.com, I obtained the information about what the top watched sports are watched as entertainment. I found out that the order from 1 - 8 is football,baseball,basketball,hockey,soccer,racing,and last is golf. My partner and I focused on the year 2015 for all the sets of data because it was the most recent information we found. The data for baseball contained information about player’s income from 1985-2015. Here (table 2)I found out that the highest income in 2015 for baseball was $32,571,000 and the lowest recorded was $507,000. The data for soccer contained information about the U.S. major soccer league and the player’s salaries in 2015. After analyzing the data(table1), I recorded that the highest income that year recorded was $6,660,000 and lowest was $50,000. Comapring the numbers side-by-side showed me that a career in major league baseball makes a significant amount more income than soccer does. About half the soccer players made $100,000 or less in 2015, while half of the baseball players made 5x more at $500,000. I also learned that there are more players in baseball than soccer. After my thorough analysis of data, I am able to answer the question ‘Does the most watched sport receive the most income?’. I have come to the conclusion that the most watched sport does receive the most income. Because baseball is the 2nd most watched sport in America and made an average of $4,313,132 in 2015, while soccer is only the 5th watched and made an average of $264,402 in 2015. The major difference in income and the higher ranking in America’s entertainment shows that the more watched sport makes a higher income. In line graph 1, it shows that soccer has a higher number of people at a lower income, while there are less people at a higher income for baseball.

1. Insert the visualizations (graphs, tables, or other resources) you created below. Title each such resource as a numbered Table or Figure and include one or more sentences referring to each Table or Figure in your explanation of the knowledge you gained in part C.

table 1: soccer table 2:baseball

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***SOCCER:*** |  |  | BASEBALL: |  |
| Income: | Amount of people: |  | Income: | Amount of people |
| 0 |  |  | 0 |  |
| 100,000 | 284 |  | 500,000 | 0 |
| 200,000 | 143 |  | 1,000,000 | 319 |
| 300,000 | 71 |  | 1,500,000 | 60 |
| 400,000 | 22 |  | 2,000,000 | 35 |
| 500,000 | 12 |  | 2,500,000 | 40 |
| 600,000 | 3 |  | 3,000,000 | 40 |
| 700,000 | 8 |  | 3,500,000 | 28 |
| 800,000 | 7 |  | 4,000,000 | 18 |
| 900,000 | 2 |  | 4,500,000 | 28 |
| 1,000,000 | 2 |  | 5,000,000 | 13 |
| 1,100,000 | 1 |  | 5,500,000 | 17 |
| 1,200,000 | 4 |  | 6,000,000 | 15 |
| 1,300,000 | 1 |  | 6,500,000 | 13 |
| 1,400,000 | 0 |  | 7,000,000 | 12 |
| 1,500,000 | 0 |  | 7,500,000 | 12 |
| 1,600,000 | 0 |  | 8,000,000 | 7 |
| 1,700,000 | 1 |  | 8,500,000 | 12 |
| 1,800,000 | 0 |  | 9,000,000 | 9 |
| 1,900,000 | 0 |  | 9,500,000 | 11 |
| 2,000,000 | 0 |  | 10,000,000 | 6 |
| 2,100,000 | 1 |  | 10,500,000 | 12 |
| 2,200,000 | 0 |  | 11,000,000 | 2 |
| 2,300,000 | 0 |  | 11,500,000 | 12 |
| 2,400,000 | 0 |  | 12,000,000 | 2 |
| 2,500,000 | 1 |  | 12,500,000 | 12 |
| 2,600,000 | 0 |  | 13,000,000 | 7 |
| 2,700,000 | 0 |  | 13,500,000 | 8 |
| 2,800,000 | 0 |  | 14,000,000 | 2 |
| 2,900,000 | 1 |  | 14,500,000 | 10 |
| 3,000,000 | 0 |  | 15,000,000 | 1 |
| 3,100,000 | 0 |  | 15,500,000 | 6 |
| 3,200,000 | 0 |  | 16,000,000 | 1 |
| 3,300,000 | 0 |  | 16,500,000 | 6 |
| 3,400,000 | 0 |  | 17,000,000 | 3 |
| 3,500,000 | 0 |  | 17,500,000 | 3 |
| 3,600,000 | 0 |  | 18,000,000 | 1 |
| 3,700,000 | 0 |  | 18,500,000 | 4 |
| 3,800,000 | 0 |  | 19,000,000 | 1 |
| 3,900,000 | 0 |  | 19,500,000 | 0 |
| 4,000,000 | 1 |  | 20,000,000 | 3 |
| 4,100,000 | 1 |  | 20,500,000 | 4 |
| 4,200,000 | 0 |  | 21,000,000 | 0 |
| 4,300,000 | 0 |  | 21,500,000 | 5 |
| 4,400,000 | 0 |  | 22,000,000 | 1 |
| 4,500,000 | 0 |  | 22,500,000 | 4 |
| 4,600,000 | 1 |  | 23,000,000 | 1 |
| 4,700,000 | 0 |  | 23,500,000 | 3 |
| 4,800,000 | 1 |  | 24,000,000 | 1 |
| 4,900,000 | 0 |  | 24,500,000 | 3 |
| 5,000,000 | 0 |  | 25,000,000 | 1 |
| 5,100,000 | 0 |  | 25,500,000 | 2 |
| 5,200,000 | 0 |  | 26,000,000 | 0 |
| 5,300,000 | 0 |  | 26,500,000 | 0 |
| 5,400,000 | 0 |  | 27,000,000 | 0 |
| 5,500,000 | 0 |  | 27,500,000 | 0 |
| 5,600,000 | 0 |  | 28,000,000 | 0 |
| 5,700,000 | 2 |  | 28,500,000 | 1 |
| 5,800,000 | 0 |  | 29,000,000 | 0 |
| 5,900,000 | 0 |  | 29,500,000 | 0 |
| 6,000,000 | 0 |  | 30,000,000 | 0 |
| 6,100,000 | 2 |  | 30,500,000 | 0 |
| 6,200,000 | 0 |  | 31,000,000 | 0 |
| 6,300,000 | 1 |  | 31,500,000 | 0 |
| 6,400,000 | 0 |  | 32,000,000 | 0 |
| 6,500,000 | 0 |  | 32,500,000 | 0 |
| 6,600,000 | 0 |  | 33,000,000 | 1 |
| 6,700,000 | 1 |  |  |  |
|  |  |  |  | Total = 818 |
|  | Total = 574 |  |  |  |

linegraph1:



1. Describe the methods that were used to analyze the data. Enough detail should be given to allow replication. Discuss the tools you used as well as the manipulations performed with your data. You might separately address the activities of identifying data, merging or cleaning data, filtering, clustering and classifying data, transforming data, and visualizing data. This written description must also address the scalability of your techniques to data sets several orders of magnitude larger.

To analyze data, my partner and I used google sheets. Setting the actual income value of soccer and baseball side-by-side allowed the analysis to be much easier. to find the average value of each sport’s income I used the equation =Average(E2:E820). The first number in the parenthesis is the first number in the table and the second is the last number as this shows the range/ length of the list of values to calculate the average for. I also founded the highest income my inputting the equation =max(E2:E20). This equation checks through all the values listed on the range given and searches for the biggest number value. I also checked for the lowest income by using the question =min(E2:E820). This checks through all the values in the range types and searched for the smallest value. Calculating the average income, highest income, and lowest income, allowed me and my partner know what number to go by in the income values. Because the range of incomes was so big, we decided to make the incomes(x) be increased by $500,000 every new line in the table. Then, we realized that the visualization would be much more understandable if it showed how many people earned from a certain range on income rather than only showing the individual sets of income. In order to find the amount of people from certain ranges on incomes, we used the equation =countifs($A$2:$A$575,">"&J4-1,$A$2:$A$575,"<"&J5). The yellow numbers/ letters show the range that the function counts through while the purple and blue values specify the income values it must be greater or less than. This set equation allows us to only type the equation a few times because after a while, when we clicked and dragged down to the empty boxes in the table, google sheets saw the pattern in the equation and finished the rest for us. This was repeated another time for the second sport.

**Scoring Rubric**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Standard** | **4** | **3** | **2** | **1** | **0** |
| ***9-12.DA.10 - Visualizations***  *Create data visualizations to help others better understand real-world phenomena.* | Insightful questions seek connections among many significant attributes. Data set(s) support a rich investigation. Consistent strategic and purposeful use of tools. Prose clearly and thoroughly explains how the data was analyzed. Meaningful artifacts clearly communicate answers to all questions. Visualizations are correct and significantly enhance the answers to the questions posed. The report clearly explains how the analysis contributes insight to the area of investigation. The entire method could easily scale to use a much larger data set. | Good questions seek connections among some significant attributes. Data set(s) support a reasonable investigation. Strategic and appropriate use of tools. Prose adequately explains how the data was analyzed with minor ambiguities or missing details. Meaningful artifacts communicate answers to most questions, with minor areas for improvement. Visualizations are mostly correct and contribute to answering the questions posed. The report clearly explains how the analysis contributes information to the area of investigation. With minor improvements in efficiency, the method could scale to use a much larger data set. | Good questions seek connections among limited attributes. Data set(s) support a limited investigation. Somewhat appropriate use of tools. Prose explains how the data was analyzed with significant ambiguities or missing details. Artifacts communicate answers to some questions with some inconsistent or inappropriate analysis. Visualizations are mostly correct and make a limited contribution to answering the questions posed. The report explains how the analysis contributes information to the area of investigation. With significant revision, the method could scale to use a much larger data set. | Questions do not connect multiple attributes. Data does not support the area of investigation. Use of tools lacks strategy or purpose. Prose predominantly is missing necessary details about how data was analyzed. Artifacts are inconsistent with data or analysis. Visualizations are incorrect or weakly connect the data and the investigation. The report confuses or misleads and weakly connects data to the area of investigation. The method could not scale to use a much larger data set without a new approach. | No evidence of ability to analyze or display data. |