DataShot: Automatic Generation of Fact Sheets from Tabular Data

Yun Wang*, Zhida Sun*, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang

APPENDIX A
EXAMPLES FOR FACTS TAXONOMY

We showcase the examples borrowed from the award-wining infographic dataset for all the 11 fact taxonomy listed in section 3.4.

Value (24.5%; 194): Retrieve the exact value of data attribute(s) under a set of specific criteria. Such facts answer the question of "what is/are the value(s) of $\{A, B...\}$ in the criteria of $\{X, Y...\}$" (Figure 1), *e.g.*, how many horses have won two out of tree Triple Crown Races?

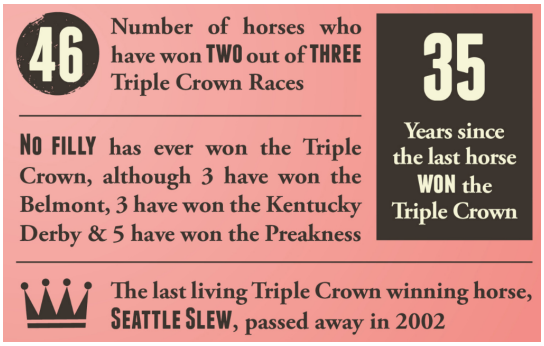

Figure 1. An example of the "value" type of fact [1].

Proportion (15.0%; 119): Measure the proportion of selected data attribute(s) within a specified set. Such facts answer the question of "what is the proportion of data attribute(s) $\{A, B...\}$ in a given set S" (Figure 2), *e.g.*, what is the percentage of protein in the diet on Sunday?

Difference (14.4%; 114): Compare any two/more data attributes or compare with previous values along with the time series. Such facts answer the question of "what is the difference between data attributes $\{A, B...\}$ within a given set S" (Figure 3), *e.g.*, what is the bed-blocking situation of a local London hospital compared with the UK average?

Distribution (11.5%; 91): Demonstrate the amount of value shared across the selected data attribute or show the breakdown of all data attributes. Such facts answer the question of "what is the summary/overall distribution over the data attribute(s) $\{A, B...\}$" (Figure 4), *e.g.*, what is the number distribution of all the unicorn companies over their age?

- *Y. Wang, H. Zhang, W. Cui, and D. Zhang are with Microsoft Research Asia. E-mails: {wangyun, haizhang, weiweicu, and dongmeiz}@microsoft.com*
- *Z. Sun, K. Xu, and X. Ma are with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. Work done during Z. Sun and K. Xu's internship at Microsoft Research Asia. E-mails: {zhida.sun, kxuak}@connect.ust.hk and mxj@cse.ust.hk*
- **These authors contributed equally to this work.*

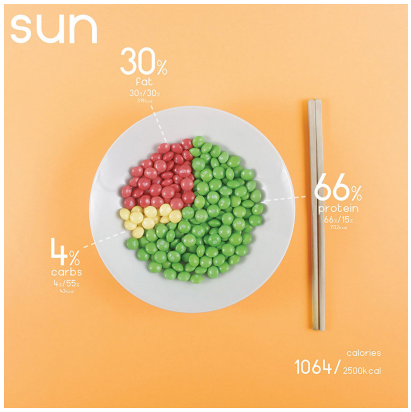

Figure 2. An example of the "proportion" type of fact [2].

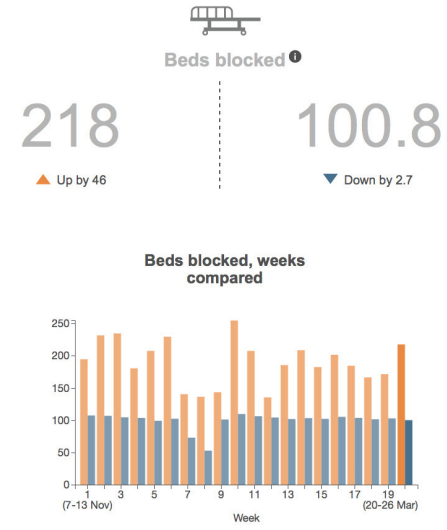

Figure 3. An example of the "difference" type of fact [3].

Trend (10.2%; 81): Present a general tendency over a time segment. Such facts answer the question of "what is the trend of the data attributes $\{A, B...\}$ over a period of time T" (Figure 5), *e.g.*, what is the budget trend for the border patrol program from 1990 to 2013?

Rank (9.1%; 72): Sort the data attributes based on their values and show the breakdown of selected data attributes. Such facts answer the question of "what is the order of the selected data attribute(s) $\{A, B...\}$" (Figure 6), *e.g.*, what are the top three reasons for consumers to engage in showrooming?

Aggregation (5.5%; 44): Calculate the descriptive statistical indicators (*e.g.*, average, sum, count, *etc.*) based on the

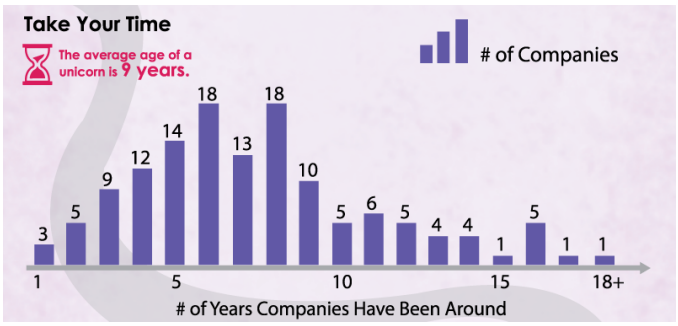

Figure 4. An example of the "distribution" type of fact [4].

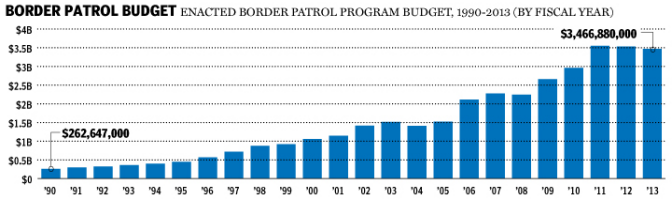

Figure 5. An example of the "trend" type of fact [5].

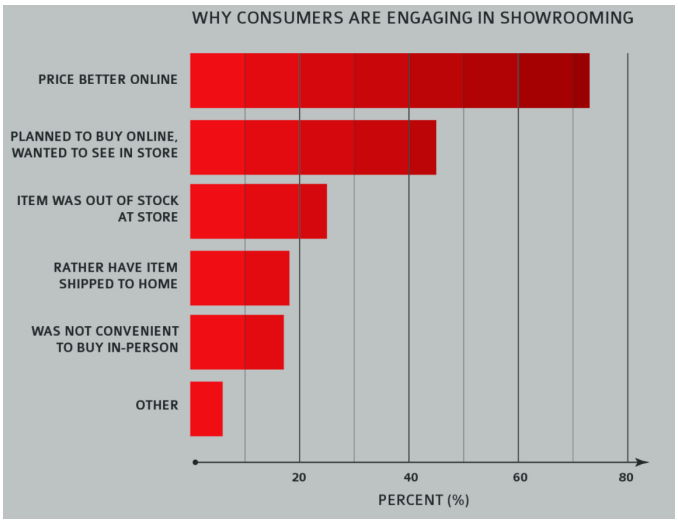

Figure 6. An example of the "rank" type of fact [6].

Figure 7. An example of the "aggregation" type of fact [7].

data attributes. Such facts answer the question of "what is the value of statistics function F over the data attribute(s) $\{A, B...\}$" (Figure 7), *e.g.*, what is the national average price for regular gas in July 2008?

Association (4.5%; 36): Identify the useful relationship between two data attributes or among multiple attributes. Such facts answer the question of "what is the correlation between/among data attributes $\{A, B...\}$ over a given set S" (Figure 8), *e.g.*, what is the relationship between products associated with the same vendor city?

Extreme (3.3%; 26): Find the extreme data cases along with the data attributes or within a certain range. Such facts answer the question of "what is/are the top/bottom N or -est value regarding attribute(s) $\{A, B...\}$" (Figure 9), *e.g.*, which character has the most epigrams in Oscar Wilde's work?

Categorization (1.4%; 11): Select the data attribute(s) that satisfy certain conditions. Such facts answer the question of "what is/are the data attribute(s) $\{A, B...\}$ which satisfy conditions $\{X, Y...\}$"(Figure 10), e.g., what are the popular names for boys in 1944 and 2004?

Outlier (.6%; 5): Explore the unexpected data attribute(s) or statistical outlier(s) from a given set. Such a fact answers the question of "what are the exceptional data attribute(s) $\{A, B...\}$ in a given set S" (Figure 11), *e.g.*, which song has the most unique words from the Beatles?

APPENDIX B
EXAMPLE FACT SHEETS FROM DATASHOT

We provide more example fact sheets generated by DataShot in Figure 12-Figure 19. Meanwhile, we also demonstrate several unsatisfied examples to show the capabilities of DataShot and identify the problems that we should improve in our future work.

1) Example fact sheets with unclear semantic meaning in the data fields (*e.g.*, the binary data case); or the semantic dependency issue as discussed in the Limitations subsection (*e.g.*, Shanghai and China should not be treated as two separate fields as they share close semantic relations). An unclear explanation example as the chart marked in the red frame in Figure 17.
2) Example fact sheets with insufficient icons/images source (Figure 17). As our local image datasets cannot cover all the topics/entities provided in the source data sheet, we designed the algorithm to pick up any available icons/images by traversing all the potential subjects in the description.
3) Example fact sheets with inefficient layout (Figure 18). The automatic layout algorithm currently optimizes fact sheet elements based on their contents and spatial relationships. However, one special case may happen when the algorithm cannot effectively utilize all the presentation space. For example, the chart in the red frame of Figure 18 shows that DataShot makes the last chart occupy the last row of the fact sheet.
4) Example fact sheets with color scheme inconsistency problem (Figure 19). Two different charts in one fact sheet share the same color encoding scheme in the marked-out area under the current design rules.

APPENDIX C
QUESTIONNAIRE DESIGN

The following questions are designed for collecting participants' experience of using DataShot.

1) The facts extracted from the data are insightful.
2) The data facts presented by DataShot are comprehensive.
3) I found the visual design in DataShot was easy to understand.

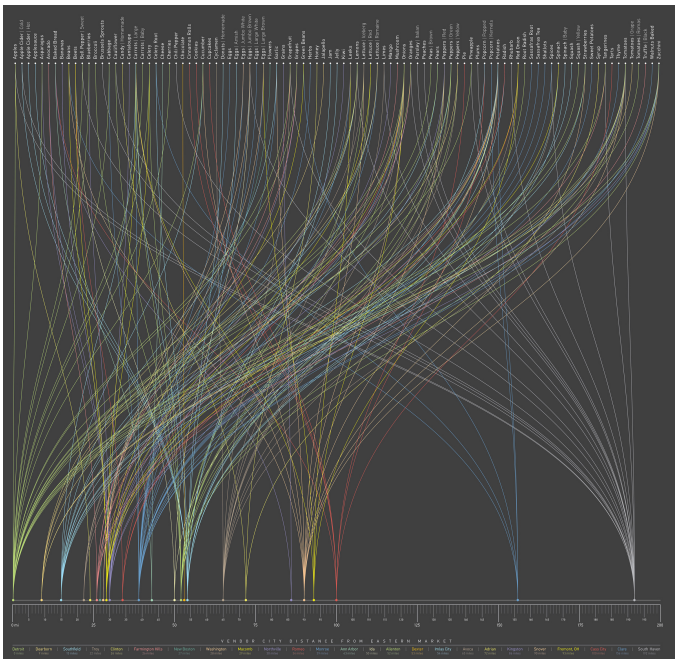

Figure 8. An example of the “association” type of fact [8].

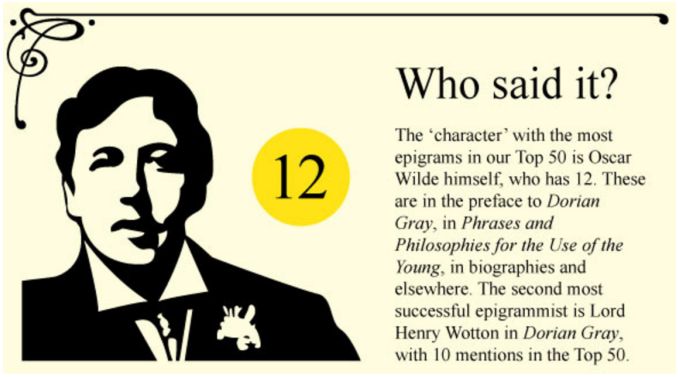

Figure 9. An example of the “extreme” type of fact [9].

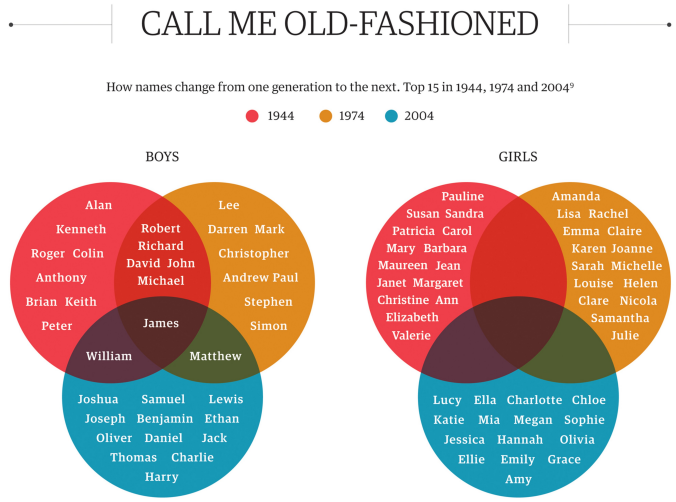

Figure 10. An example of the “categorization” type of fact [10].

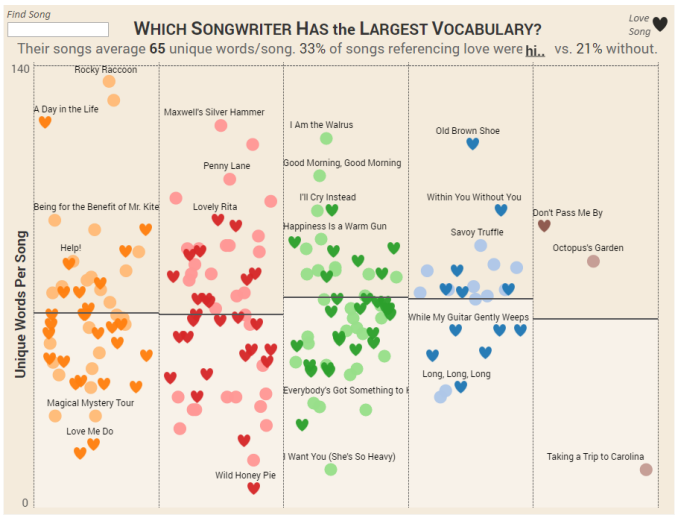

Figure 11. An example of the “outlier” type of fact [11].

4) DataShot was effective for providing facts from tabular data.
5) I thought the visual design was aesthetically appealing.
6) DataShot could provide expressive visual presentations.
7) DataShot was useful for presenting interesting facts extracted from the tabular data.
8) DataShot was easy for me to get the infographic fact sheet from the tabular data.

REFERENCES

[1] Activ8Social, “A race against history information is beautiful awards.” https://www.informationisbeautifulawards.com/showcase/371-a-race-against-history. (Accessed on 06/21/2019).
[2] R. MacEachern, “Design x food information is beautiful awards.” https://www.informationisbeautifulawards.com/showcase/67-design-x-food. (Accessed on 06/21/2019).
[3] J. W. C. T. R. M. C. A. Branwen Jeffreys, Dominic Bailey, “Nhs winter: Weekly a & e tracker information is beautiful awards.” https://www.informationisbeautifulawards.com/showcase/581-nhs-winter-weekly-a-e-tracker. (Accessed on 06/21/2019).
[4] S. insights, “The quest to find a unicorn: Vc myth becomes investment payoff in 2015 information is beautiful awards.” https://www.informationisbeautifulawards.com/showcase/1093-the-quest-to-find-a-unicorn-vc-myth-becomes-investment-payoff-in-2015. (Accessed on 06/21/2019).
[5] M. Faille, “Us border crisis information is beautiful awards.” https://www.informationisbeautifulawards.com/showcase/751-us-border-crisis. (Accessed on 06/21/2019).
[6] B. M. Cristina Vanko, “Showrooming information is beautiful awards.” https://www.informationisbeautifulawards.com/showcase/120-showrooming. (Accessed on 06/21/2019).
[7] H. Jones, “The math of cheap oil information is beautiful awards.” https://www.informationisbeautifulawards.com/showcase/906-the-math-of-cheap-oil. (Accessed on 06/21/2019).
[8] J. Hagen, “Eastern market double sided poster information is beautiful awards.” https://www.informationisbeautifulawards.com/showcase/34-eastern-market-double-sided-poster. (Accessed on 06/21/2019).
[9] Z. V. Adam Frost, “And the oscar goes to... information is beautiful awards.” https://www.informationisbeautifulawards.com/showcase/78-and-the-oscar-goes-to. (Accessed on 06/21/2019).
[10] J. Lenman, “What’s in a name? information is beautiful awards.” https://www.informationisbeautifulawards.com/showcase/116-what-s-in-a-name. (Accessed on 06/21/2019).
[11] A. E. McCann, “Beatles analysis information is beautiful awards.” https://www.informationisbeautifulawards.com/showcase/1146-beatles-analysis. (Accessed on 06/21/2019).

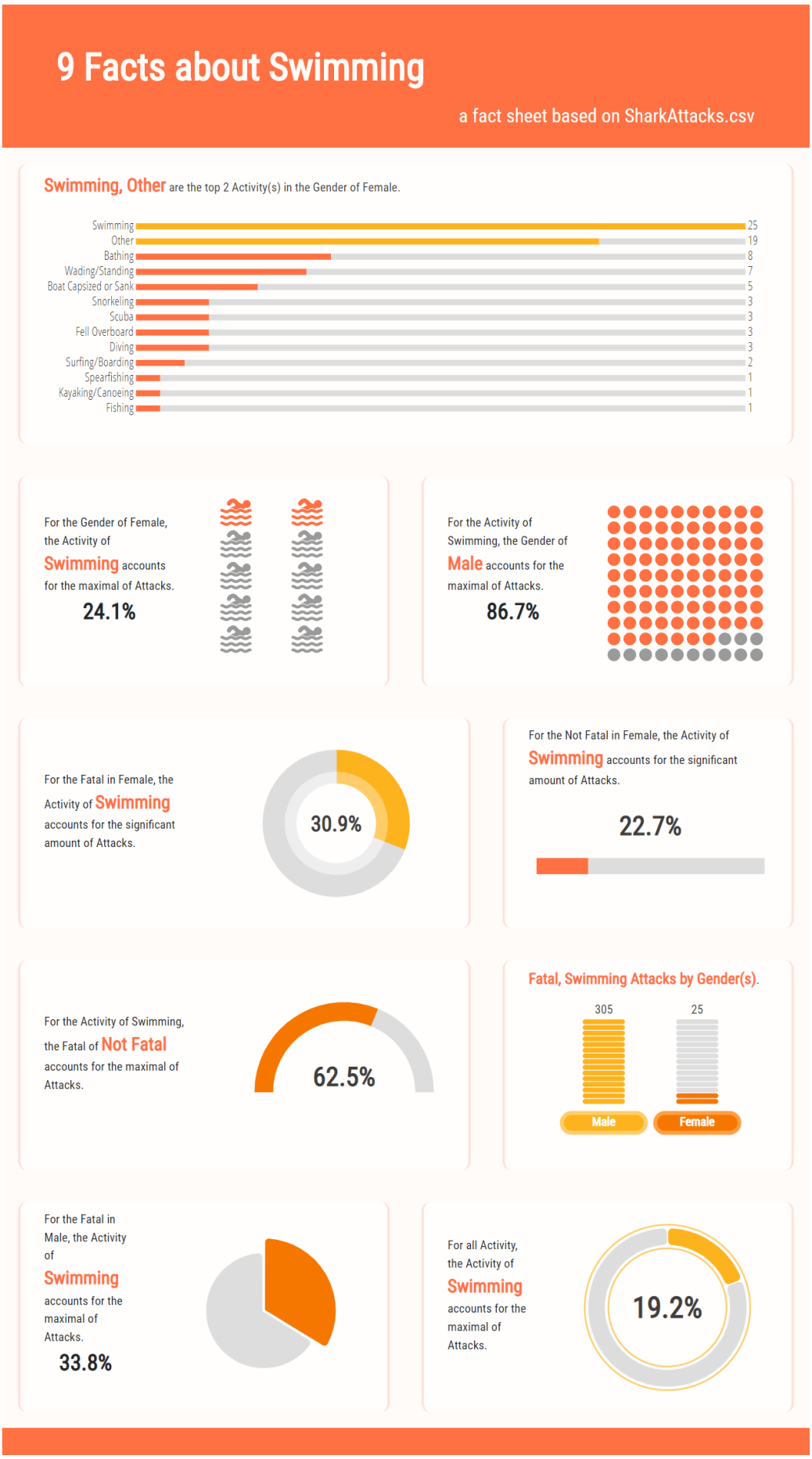

Figure 12. An example fact sheet of swimming event based on SharkAttacks.csv.

Figure 13. An example fact sheet of sports car sales based on CarSales.csv.

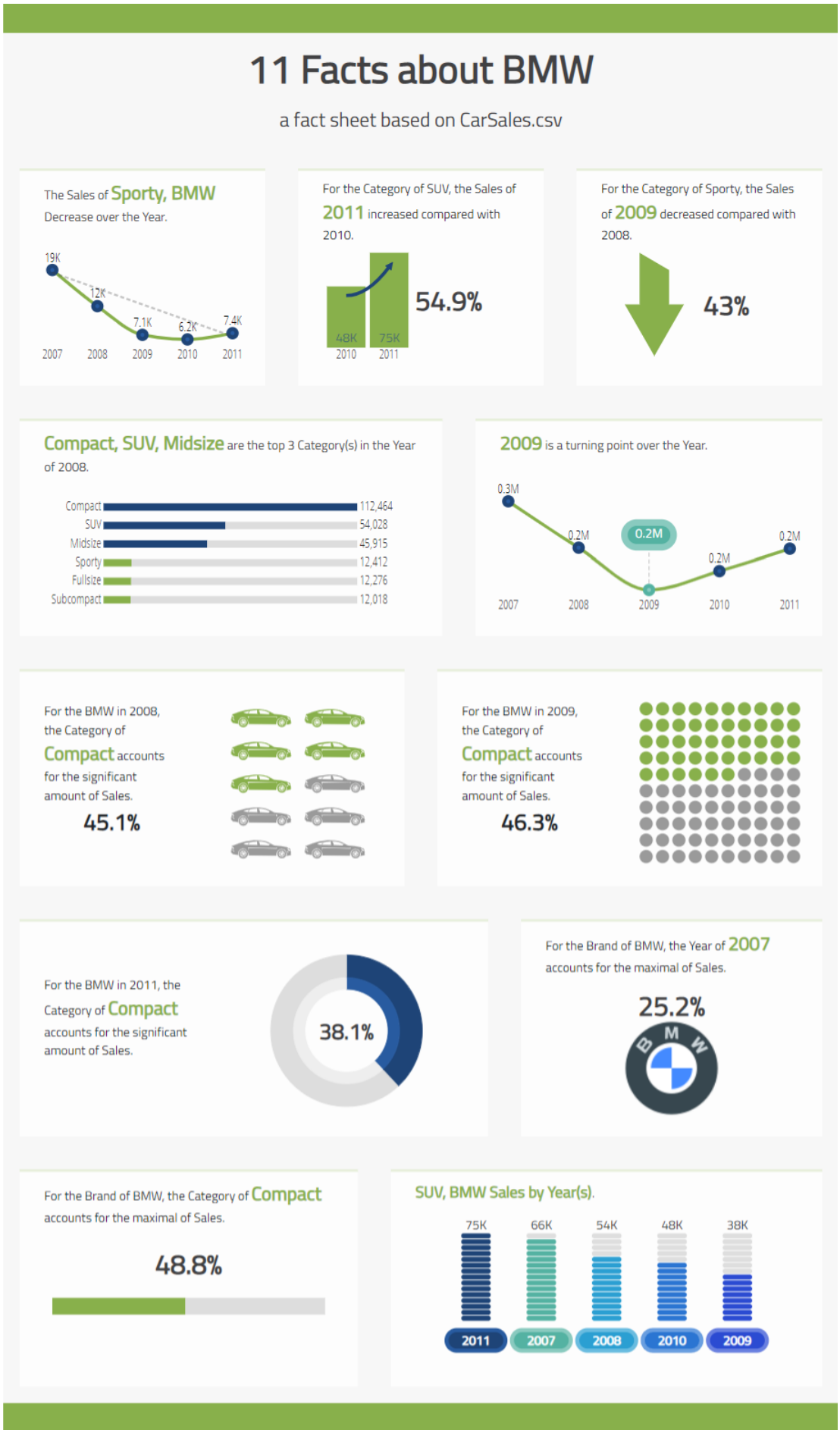

Figure 14. An example fact sheet of BMW sales based on CarSales.csv.

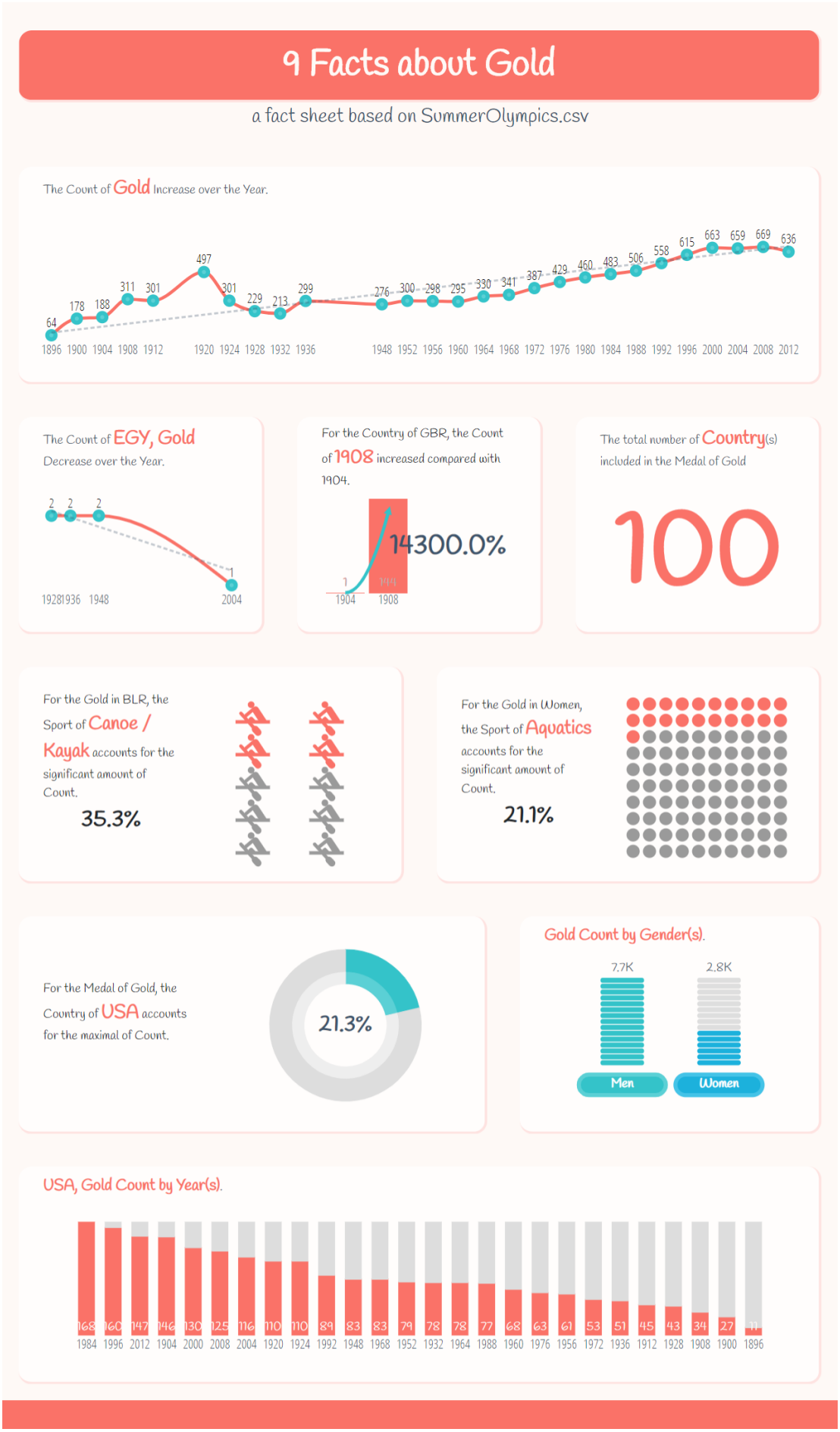

Figure 15. An example fact sheet of golden award winning event based on SummerOlympics.csv.

Figure 16. An example fact sheet of fatal event based on SharkAttacks.csv.

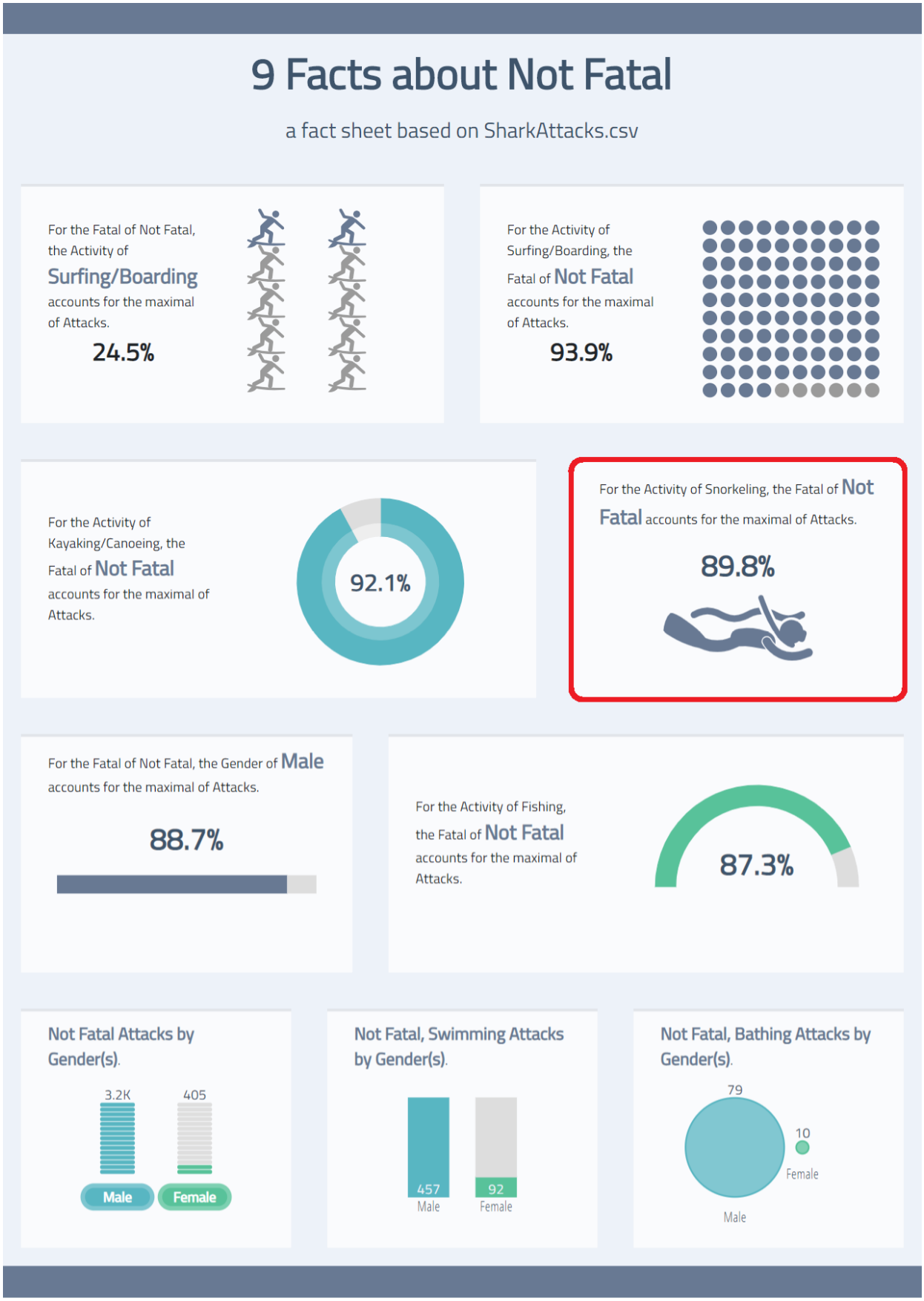

Figure 17. An example fact sheet of not fatal event based on SharkAttacks.csv.

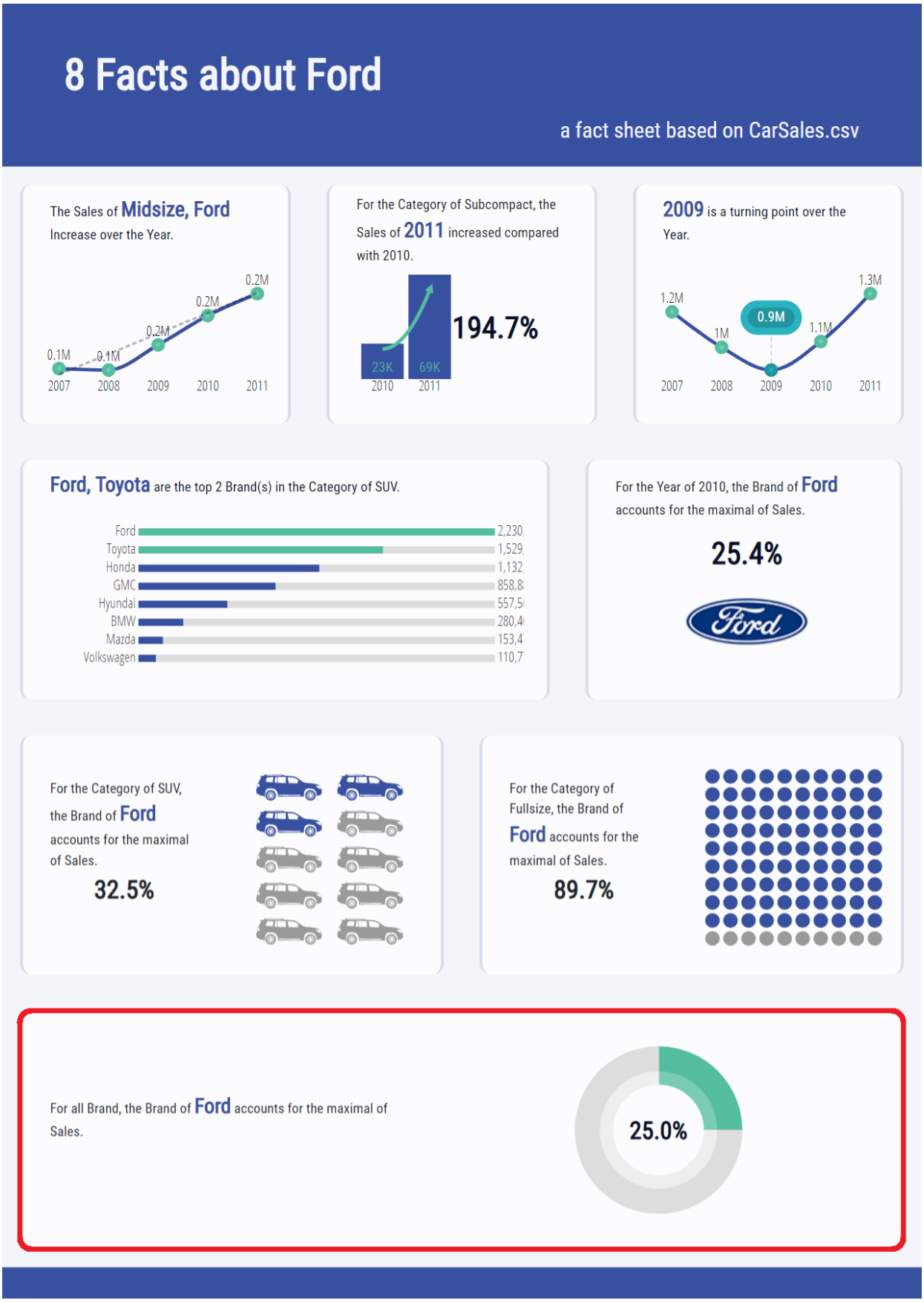

Figure 18. An example fact sheet of Ford sales based on CarSales.csv.

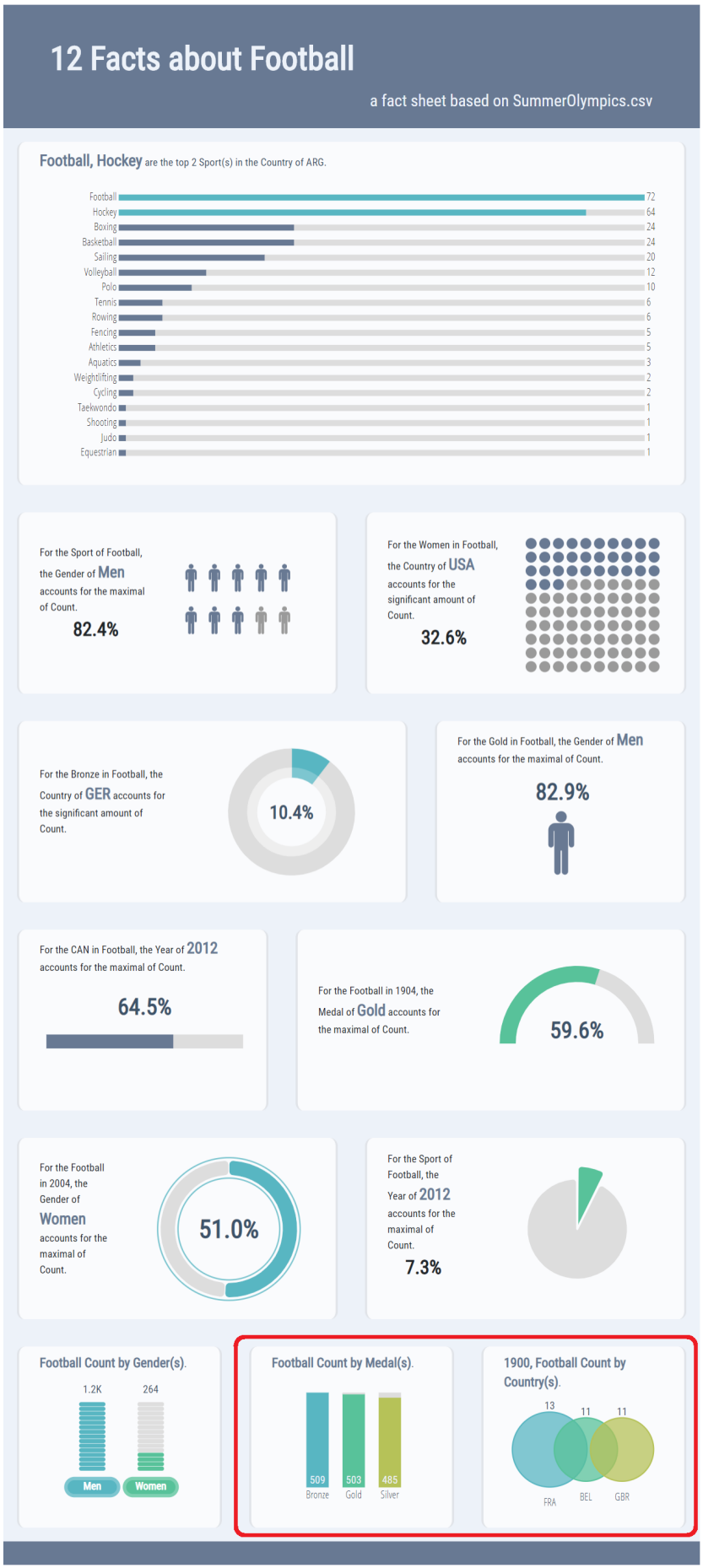

Figure 19. An example fact sheet of football event based on SummerOlympics.csv.