Have you had an accident at work?

Data Visualisation: Harry Talbot

hwt1g15

I. Data Story Summary

The reality of accidents at work is not often well understood, as (thankfully) due to modern health and safety measures many people work in very safe environments with little or no exposure to any life-threatening risk. The data story aims to draw attention to the reality of accidents in the workplace - particularly fatal - focusing on the variation in rates and causes between occupations, industries, employment location and employee characteristic. The story adopts a linear path that is navigated with the arrow keys in a page-like manor – rather than scrolling – so each visualisation is clearly defined; however, the reader can explore data in more detail at each visualisation and move forward and backward as they please.

II. Dataset Summary

I decided early on that I would focus on data from one country, and tie in small facts from around the world where possible. Similarly, when deciding on a timeframe, I decided to take focus on one year, and tie in small pieces of data from other years.

So that the story is interesting and relevant to the audience, I initially considered data from the U.K. and the U.S.A. as the social and economic aspects of the two are similar. U.K. data proved to be difficult to visualise because:

- **The data is not grouped beyond industry level:** I would not be able to look at specific jobs, as cases were pre-grouped into vague industries such as 'Development of Building Projects', and
- **The case counts are much lower:** with a population a fifth of the size, the number of fatal injuries per industry level rarely exceeded 15-20, and so without more data I could not be confident in the accuracy or the rates I was presenting.

Alternatively, American data uses Standard Occupation (SOC) codes to group much more finely (filtering down to occupations such as 'Supervisors of Sales Workers'). The American data was sourced from the Bureau of Labor Statistics (BLS) [1], specifically the programs

- Injuries, Illnesses, and Fatalities (IIF)
- Employment Projections (EMP)
- Current Population Survey (CPS)
- Occupational Employment Statistics (OES)

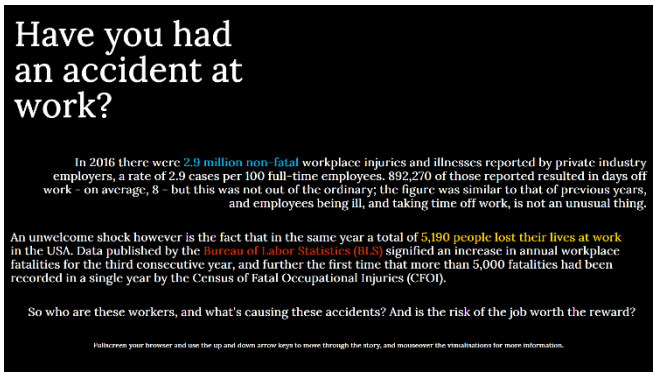
Figures for occupational employment were taken from the EMP program, as those from the CPS and OES did not include self-employed workers and were not consistent with each other (whereas the employment figures for 2016 in the EMP 2016-2026 prediction did include self-employed workers). The 2016-2026 prediction is the only one currently available, else more recent data would have been used (such as a 2018-2028 prediction).

The IIF's 2016 'Census of Fatal Occupational Injuries & Occupation' and the 'Case and Demographic Characteristics for Work-related Injuries and Illnesses' detailed fatalities and injuries respectively by cause and occupation and provided all the accident data needed for the visualisations. The data came in the form of a set of XLSX tables, one per occupation group, that were then scraped together into one file.

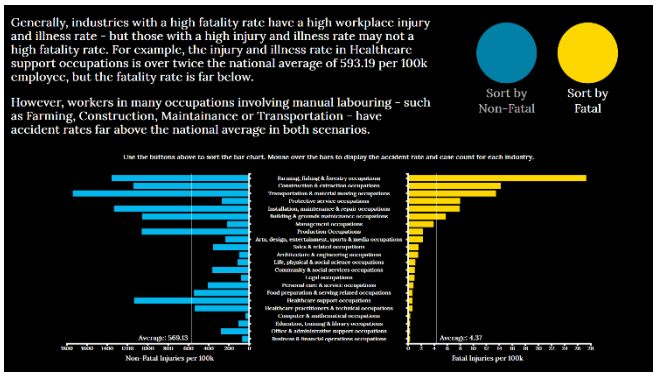
The employment, injury and fatality data were combined in Tableau to eliminate duplication, and the resulting table exported to a CSV file. Blank values (~50, mainly for salary) were filled in manually from data elsewhere via the BLS.

III. Visualisations

An initial rhetorical question and text introduces the topic and explains the nature of the story, and hints at the takeaway message.



A. *Fatal vs. Non Fatal Accidents per 100k by industry: Back-to-Back horizontal bar.*



1) *Description*

Depicting the Fatal and Non-Fatal accident rate per 100k employees for broad occupation codes, this visualisation

compares the two rates on a per-industry basis. The reader can sort the chart so that either side is in descending order, with the opposing side animating into position with it. When moving the cursor over a specific bar, numerical labels at either end become visible to give exact figures for that industry. Supporting highlights noticeable trends and guides the view of the reader.

2) Justification

As well as being a straightforward opening visualization, the reader should immediately notice the trend to the bars rather than the actual rates themselves, which are otherwise difficult to compare as they are of significantly different magnitudes. The scales are clear, as so to avoid misleading the reader into thinking the results are directly comparable (and that there are relatively similar numbers of fatal and non-fatal accidents)

3) Narrative Design Patterns

The comparative nature of this visualisation is ideal for observing the differences between the two data sets at a high level. The scales mean that the two rates can be compared visually, where usually the magnitudes are far different.

4) Strengths and Weaknesses

The trends and lack of correlation between the two categories is clear to see, and the average bars further contextualise how unusual some of the rates are. Colours from the introduction are carried through for consistency.

5) Improvements

Although this is just an introductory visualisation it would be useful for it to show more contextual information on mouseover, such as specific comparisons and observations per industry. The ability to change the granularity of the chart – such as grouping by industry, broad or specific occupation – would allow for comparisons in the trends to take place at different levels, rather than just by industry.

B. *Fatal vs. Non Fatal Accidents per 100k by Cause: Stacked Percentage Contribution*

1) Description

This visualisation illustrates how different causes make up the total number of fatal and non-fatal accidents at work by proportion, and introduces the key causes that data is grouped into throughout the story. Moving the cursor over the bars displays the percentage that bar makes up, and supporting texts make observations about the data, colouring texts where possible to link the series in the bars to the commentary. At the end, the visualisation poses the question 'So what are the deadliest jobs in America today?' before leading into visualisation 3.

2) Justification

As the values of the series (here the causes) differ largely between the bars it is easy to make rough comparisons, before looking at the actual percentages when using the cursor. As I wanted to include statistics in the texts that would not need their own chart, the stacked bar was ideal as it shows multiple categories in a compact space, leaving more room for the text.

3) Narrative Design Patterns

Similar to the previous visualisation, a comparative narrative makes reference to the stacked bar and comments on the data.

4) Strengths and Weaknesses

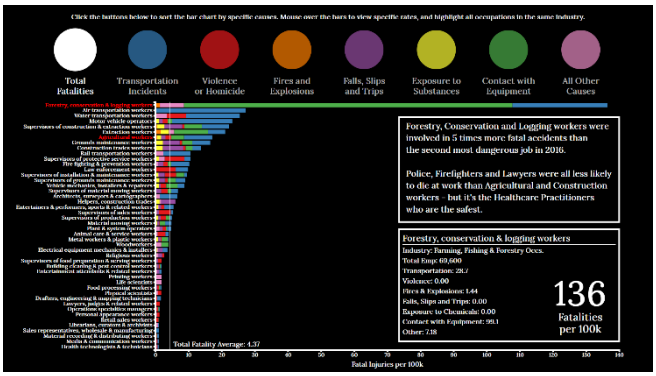
Stacked bar charts can be difficult to use for relationship analysis. As the common baseline lies on the left of the chart, it can be hard to reliably compare values other than the leftmost series or the total of the bar.

5) Improvements

Available injury data from the BLS subdivides the causes of injury further – for example, dividing transportation incidents into those involving another car, a roadside object or when the victim was a pedestrian – but the fatality data does not. I would have liked to have used a grouped vertical bar chart to show the distribution in these causes, but decided that the without fatality data as well, a non-fatal chart would not add relevant value alone in a comparative visualisation such as this. As well as this, the story focusses around fatality data predominantly after this visualisation.

Had grouped fatality data been available, then a back-to-back grouped bar could have been shown, which would make the values of the causes easier to identify. A stacked bar would not be appropriate as there would be too make categories.

C. *Fatal Accidents by Cause: Stacked Bar Chart*



1) *Description*

The third visualisation shows a sortable stacked bar chart, displaying fatality rates for specific occupations broken down by cause. The chart can be sorted by total fatality rate per 100k, or by a fatality rate for a specific cause. Depending on the sort, supporting text guides the view of the reader, comparing different roles to highlight chart features that may be unexpected. When moving the cursor over the bars, per-occupation information is shown, and axis labels for occupations in the same industry are highlighted.

2) *Justification*

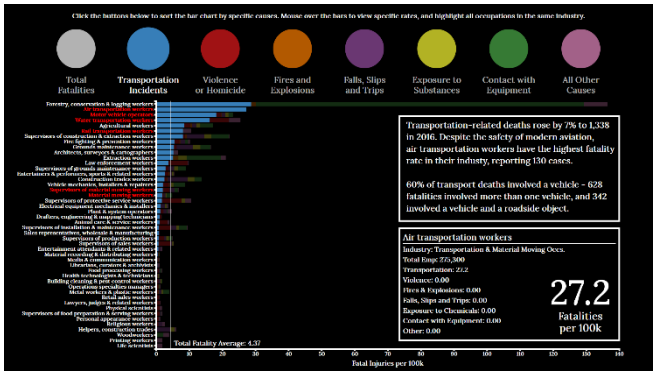
A stacked bar is an effective way to show a group's structure – in this case, the accident causes - in a set of data, while still showing the differences between the totals of the group. The colours of the causes allow for comparison between individual causes visually – when sorted, the respective bars stack against the axis so that the descending order can be seen clearly, with the rest of the bar faded; the faded parts of the stack still give some indication to the group's overall value, highlighting a part to whole relationship in the data and letting the reader continue to make comparisons through the visualisation. The total values cover a significant enough range that the stacked bar chart shows the differences well, whereas a grouped bar would have reduced the overall bar length range and made differences harder to see without value labels.

3) *Narrative Design Patterns*

Again comparative, but in a different format this visualisation breaks down the data for comparison. The visualisation is somewhat explorative, and the reader can view information in as little or as much detail as they would like.

4) *Strengths and Weaknesses*

As mentioned before, stacked charts are susceptible to being difficult to compare for series that are not aligned against the axis. To overcome this, the chart animates when sorted to position the selected series against the axis, and sorts by the series rather than the overall chart; the other series are faded but still visible, so total fatality rates can still be observed. This is shown below when sorting by transportation incidents:



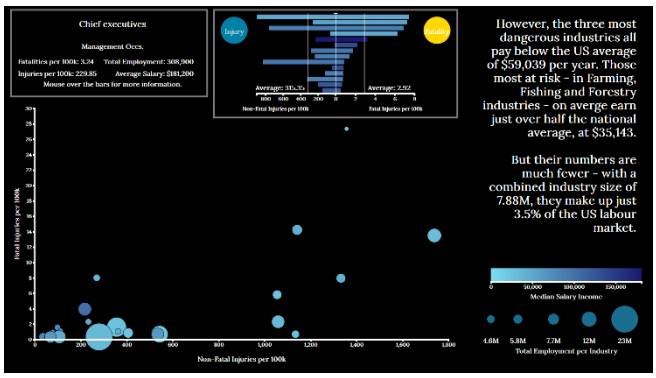
The colours of each cause are also the same as those used earlier to avoid making unintended connections or confusion.

It can be difficult to compare a bar near the top of the chart with the lower axis, so instead the fatality rate is displayed larger than other values when the cursor is moved over the bars. Other options involved either duplicating the axis along the top of the chart, or just moving it – but this either takes up too much space or simply reverses the problem.

5) *Improvements*

An option to draw the same chart with non-fatal data or perform some sort of comparison would improve the visualisation, and filters to reduce the number of bars may improve the readability. Given more time, the chart could initially show by industry, but expand industries into occupations when they are selected. The bars in the chart referenced in the text could be highlighted to improve the association for the reader.

D. *Industrial and Occupational Fatality rates by Salary and Employment Level: Scatter Plot*



1) *Description*

The scatter plots industries with their fatality and injury rate, but also encodes employment and average salary measures. Moving the cursor over plots displays a small versus-style bar for the occupations it represents, and further moving over the bars gives more details on the occupations. The act of transitioning to the next visualisation first updates the axis to include outliers, with the text updating too.

2) Justification

Scatter charts are primarily used for visualising correlation, and here it is used to show the pattern between increasing injury and fatality rate, decreasing salary and decreasing employment level. The reader initially just sees a handful of industries, that are well paid and have low injury rates. The scale then changes to include all industries, and those with higher rates are highlighted as outliers. The scale of the plot point illustrates well the change in employment size.

3) Narrative Design Patterns

The design patterns used here are a combination of an exploration with a gradual reveal. Slowly revealing the whole axis spread and occupation data helps the reader understand how the data is broken down, and the exploration elements increases active engagement as the reader can interact as much as they want.

4) Strengths and Weaknesses

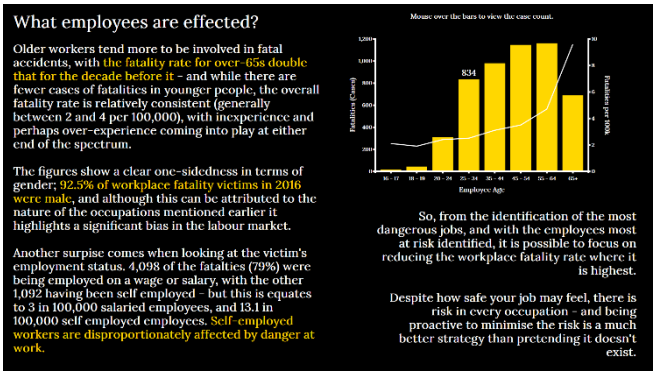
The visualisation presents a lot of information, with a logical order to filtering the data to see specific values. However, the correlation is not directly highlighted to the reader, and is only mentioned through the supporting text.

The colour scale works well for the plots, but as the same scale is used throughout the whole visualisation, occupational salaries in the versus bar can appear relatively similar when the difference can be $10-15K.

5) Improvements

Two different colour scales could be used for the plot and the versus bar, so salaries for occupations are compared relative to their industry. An average point plot may be useful for reader reference

E. Concluding Visualisation and Bar Chart



1) Description

The final visualisation uses a dual-axis bar chart to show the age of employees at death, with the number of cases illustrated by the bars and the rate by the line. The supporting text summaries other figures and concludes the story.

2) Justification

The purpose of the final visualisation was to tie in some key figures about employee demographics with a short conclusive paragraph to end the story, relating it back to the reader and giving them a conclusive message. The included chart emphasises that while there are fewer cases of fatality in older people, the rate is much higher.

3) Narrative Design Patterns

This final visualisation concludes the story with a call to action, stating that no matter how safe a job me be there is no substitute for proactively reducing the chance of an accident at work.

4) Strengths and Weaknesses

The 3 main human traits in the data – gender, age and employment category – are all summarised briefly for the reader, with key figures relating to employee characteristics highlighted in the same colour that has been used for fatality throughout the story large amount of text on this final screen means it can be hard for the reader to know where to look first

5) Improvements

In future employee characteristics could be explored more deeply, for example looking at ethnicity and linking it with location data if available. The concluding section could then be separated out from the analysis, to give a greater fell of the data story ending.

IV. Conclusion

I had intended to use a Choropleth map to look at data on a per state level, and then make comparisons to the rest of the world. Due to variation in how the data is collected between states (a combination of an old and new categorising system means any grouping I did may be inaccurate, and unlike how it is grouped in national data) and time constraints I was unable to include this visualisation. Using data from other countries such as the UK would provide more context to the reader and so given more time this is something I would include.

Similarly, more specific case data about the demographics that would have allowed me to calculate rates for a person with a combination of traits (e.g. a male firefighter aged 25-34, rather than those traits individually) was not available

through the BLS[1]. With this data the story could have taken a more 'how likely are you' approach, with the reader picking traits and the visualisation updating with various metrics.

The visualisation explores the available data in a format that is accessible and interactive to the user and answers the core questions throughout the coursework: those I had when I first started research, when I gave my presentation, and those presented at the start of the final story. Given more time, there is room for improvement though adding more interactivity or location-based data, be it from multiple states or countries.

V. References

[1] "U.S. Bureau of Labor Statistics," 2019. [Online]. Available: BLS.gov.

[1] After emailing the BLS directly to enquire about the availability of data, they responded by saying that the way the data had been collected (or at least is accessible to them now) meant they cannot compute such metrics, they are unable to release incident data by cases (like is done in the UK) for me to produce the figures myself.