InfoVis Data Visualization Project

Andrea Alfieri - 5128315
A.Alfieri-1@student.tudelft.nl
TU Delft
Delft, NL

Reinier Koops - 4704312
R.W.Koops@student.tudelft.nl
TU Delft
Delft, NL

Aditya Kunar - 5074274
A.Kunar@student.tudelft.nl
TU Delft
Delft, NL

ABSTRACT

Data visualization gives way to clear, concise and efficient visual communication. It's primary purpose is to leverage the incredibly sophisticated human perceptual system to be able to derive unique insights in order to aid the user in carrying out tasks more effectively. In that regard, our work focuses on information visualization.[1] This report subsumes our compilation of many different visualizations each defined on the basis of their respective task abstractions in addition to the visual and interactive idioms.

ACM Reference Format:
Andrea Alfieri - 5128315, Reinier Koops - 4704312, and Aditya Kunar - 5074274. 2019. InfoVis Data Visualization Project. In *Delft '19: Data Visualization Final Report, January 08, 2019, Delft, NL.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The main objective of this report is to present our work in a clear and concise manner. Thus, we have divided the report in four sections. Section 2 describes the datasets within the scope of this project and why we chose them. Section 3 gives an initial and more conversational description of the visualizations that we created in order to give the reader a general idea of our work. Section 4 elucidates our work in terms of different *task abstractions* (following the structure given by [6]). It also sheds some light on the interactive idioms being deployed for each task as well. Section 5 is used to elaborate on our choices for using various visual encodings from a purely theoretical point of view.

2 DATA

For this project, we chose the dataset released by the UK Police [5] containing information about criminal events and stop & search statistics of the UK. This allowed us to perform a diverse array of complex visualization tasks because of the proliferation of different categorical and quantitative attributes as well as the combination of spatial and temporal attribute relations. An example of these attributes are location (hierarchical attribute with granularity county/city/local authority), date (hierarchical attribute with granularity year/month/day), type of crime, outcome, gender, ethnicity, age and object of search(categorical attributes). To add more complexity we added demographic statistics such as population density, average annual income per household and unemployment rate(quantitative & sequential attributes) from UK's Office for National Statistics [3]. This data could be linked spatially via local authority codes in order to show more interesting derived attributes (e.g. crime *per capita* instead of just simply number of records). All datasets used are tabular and static in nature.

[1] The data-sets for this project include the "Police UK" dataset comprising of "crime" & "stop and search" data augmented with the demographic data for the UK as well.

Delft '19, January 08, 2019, Delft, NL

ACM ISBN 978-1-4503-9999-9/18/06...$15.00
https://doi.org/10.1145/1122445.1122456

3 THE ANALYSIS

One of our aims for this project has been to simulate the work of a data analyst. Therefore, we decided to explore the data by starting from a more general analysis, followed by more detailed visualizations. For the same reason, we first considered each dataset on its own, and only later used them in combination to create more complex visualizations.

3.1 Demographic data

We first created basic visualizations about individual attributes: population density and average annual income per household (Figure 1 and 2 respectively). This helped us to look at general distributions and outliers. A similar visualization to figure 2 has been made for unemployment and is shown in the screencast.

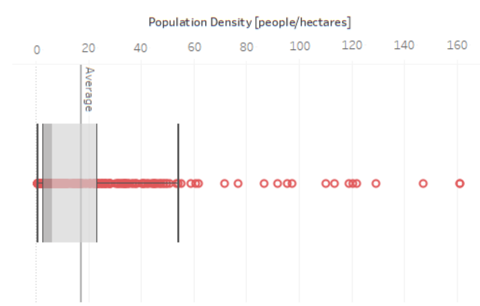

Figure 1: Box plot of population density across England

Looking at the box plot in figure 1, we immediately noticed many outliers and so we decided to learn more by showing them through a geographical point map visualization. In fact, figure 3 shows the plot of the 25 most densely populated areas and their associated characteristics on a map.

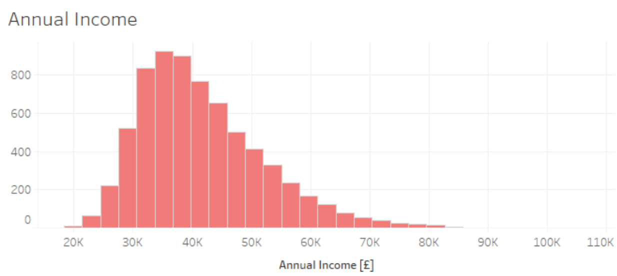

Figure 2: Histogram of income in England

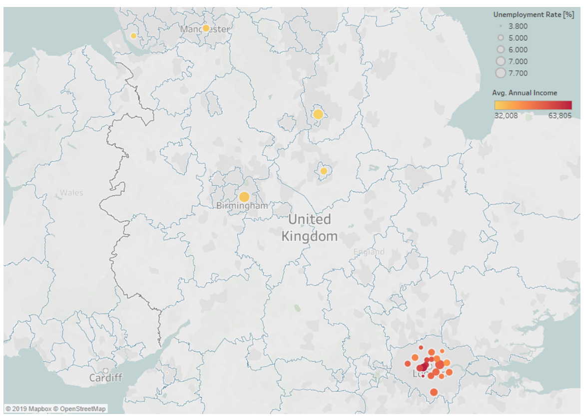

Figure 3: 25 most populated areas of England

3.2 Stop & Search Data

Because of the greater complexity and richness of information in this dataset, we decided to work on visualizations involving the use of greater interactivity between the dataset and the user to allow for a more flexible exploration of the dataset.

The task shown in figure 4 provides information about the number of searches (normalized by the population of the area), gender, ethnicity, object of search and the outcome associated with it. However, as compared to the previous visualizations, not only can the user look at multiple different views, the user can also dynamically alter them based on his/her preferences by interacting with them too. Therefore this one visualization offers a data analyst the ability to investigate freely on an encapsulation of the entire information contained in the dataset.

Interest in the evolution of the attributes over time can then be satisfied by the next visualization, shown in figure 5. Here, the user can similarly interact with the views as well.

The final visualisation for this dataset was made with the purpose of reducing the cognitive load of the user to remember and retrieve all the details about a selected location. As we know that eyes beat memory, in figure 6 the user is provided with the relevant information juxtaposed for easy comparison.

A strategy that we applied for this dataset was to use the same map layout as an interactive filter to be used throughout the three visualizations, to let the user easily adjust between them. The map also has the function of showing the search per capita of each area.

3.3 Crime Data

For this dataset, we first analyzed it separately and then contextualized it via the use of demographic data. Similar to our approach in stop and search, we included a high degree of user interaction.

Figure 7 shows a visualization in which the user can select locations on the map and view the distribution of crimes types. The map also has the function of showing the crime per capita of each area. Because the data is more geographically precise than the one used previously, it lends to the possibility of analyzing details about different areas of the same city via the use of local authority codes. In fact, in the following example provided, we are only showcasing the areas of London which are north of the Thames.

On the other hand, figure 8 already shows the interaction between demographic and crime data and tries to find interesting trends that might relate them. Because of the great quantity of data points, we decided to plot the density of the data points.

The last visualization of the project is shown in figure 9 and allows the user to look at trends over time for different crime types through the use of a streamgraph. The example shows the top 5 crime types for the coastal area in the north-east of the UK.

In conclusion, we have presented a progressive analysis to reflect the work of a data analyst utilizing various visualization techniques and interactive features so that the user can investigate even more trends than those we found interesting and decided to show in the report.

4 TASK ABSTRACTION AND INTERACTIVE IDIOMS

In this section, we provide a more standard description of our tasks as per the design space dimensions described in the paper by Schulz et al [6] and the interactive idioms being used as well. Each task, which in our case corresponds to a single visualization created, is explained in terms of *goal*, *means*, *characteristics*, *target* and *cardinality*.

4.1 Demographics - task 1

This subsection corresponds to the tasks described by figure 1 and 2.

- **Goal:** The goal is to do an *exploratory analysis* via an *undirected* search.
- **Means:** The goal is reached by performing a search on the individual attributes at the granularity for each location i.e *navigation.*
- **Characteristics:** The *high level characteristics* include looking at outliers and at the general distribution whereas the *low level characteristics* are simply the values of each data point.
- **Target:** The target attributes are population density and average annual income per household for figure 1 and 2 respectively.
- **Cardinality:***All instances* are shown to give the viewer an overview.
- **How is it shown:**A box plot to show population density and a histogram for the average annual income per household.

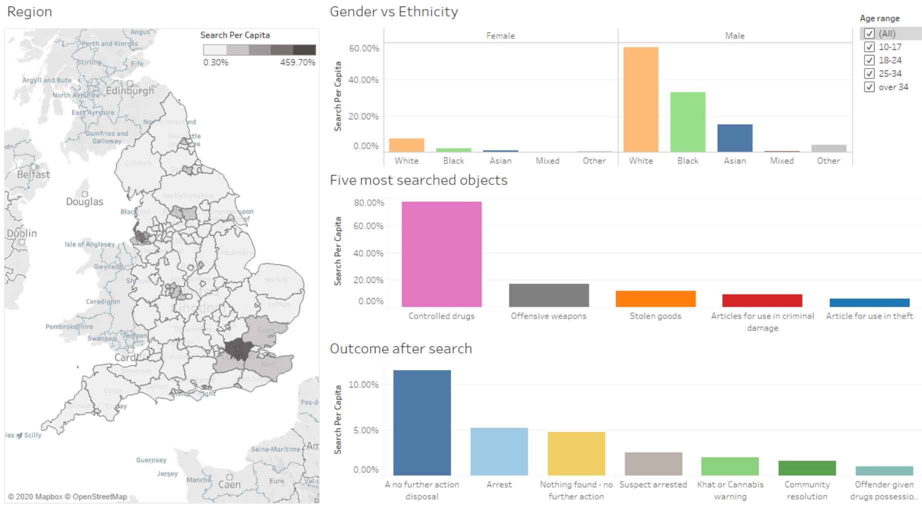

Figure 4: General info about stop and searches

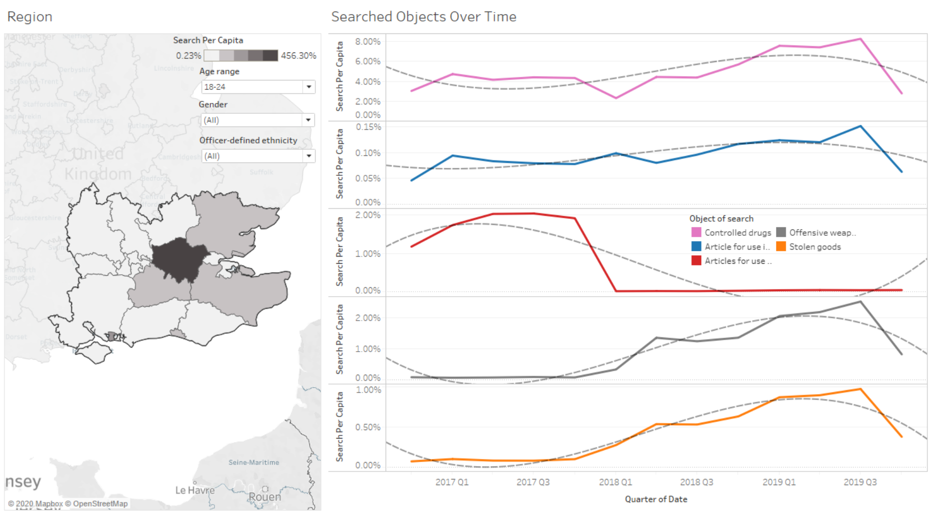

Figure 5: Trends over time for London and its surrounding provinces

4.2 Demographics - task 2

This subsection corresponds to the task described by figure 3.

- **Goal:** The goal is to run an exploratory analysis on the associated attributes for the outliers in the population density.
- **Means:** The goal is reached by performing a search or a browse on the associated attributes such as average annual income per household and unemployment rate corresponding to locations with the highest population density (i.e *navigation* and *relation*) and *reorganization* of the data by filtering to depict only the top N data points with respect to population density as customized by the user.
- **Characteristics:** This visualization highlights the interesting *low* level characteristics of the other associated attributes for the outliers with respect to population density.
- **Target:** The information is shown through *spatial relations* wherein all the above mentioned attributes are linked to geographical areas on the map.
- **Cardinality:** *Multiple instances* are shown in order to give the user a comparative viewpoint thereby putting the data in context.
- **How is it shown:** Geographic point map.

4.3 Stop & Search - Task 1

This task corresponds to figure 4.

- **Goal:** The goal of this task is to allow the user to generate a hypothesis and is therefore an *exploratory analysis*.
- **Means:** The goal is reached through *navigation*, *reorganization* and *relation* by means of making use of multiple linked

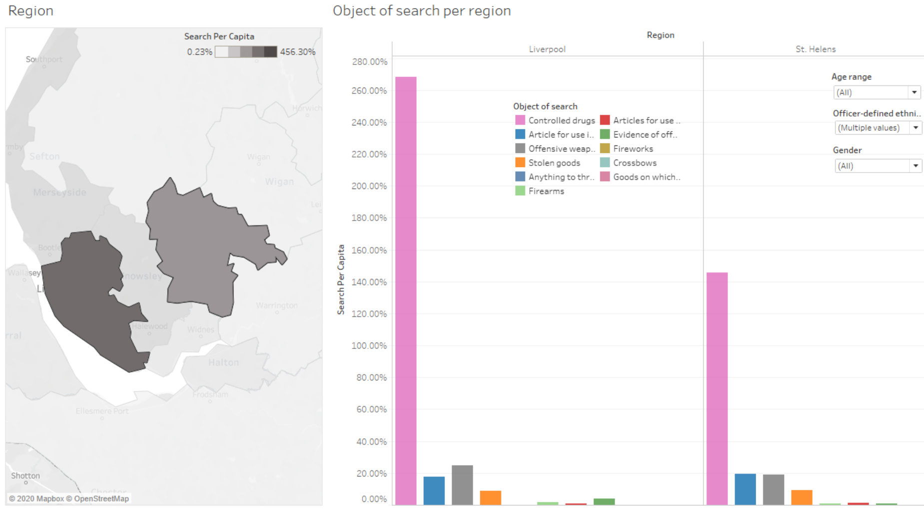

Figure 6: Comparison of regional distribution for the objects of search

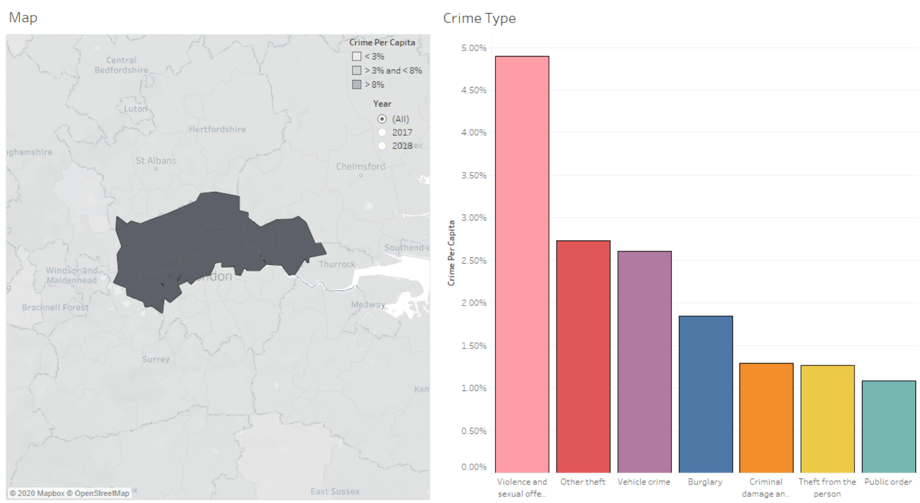

Figure 7: England crime type histogram, North London

views wherein the user can filter interactively based on multiple attributes such as locations on the map, gender, ethnicity, age and object of search.

- **Characteristics:** The task reveals *high-level data characteristics* such as outliers and general distributions for the above mentioned attributes. The user is also able to compare the *high-level data characteristics* between different categorical attributes all at once.
- **Target:** The target for this visualization are the following attributes: age, gender, ethnicity, object of search, outcome and location. Thus the task is carried out on *spatial* and *structural* relations.
- **Cardinality:** *Multiple instances* are addressed for each of the attributes in order to put the data in context.
- **How is it shown:** The first view in figure 4 corresponds to a choropleth map, the second view is a multi-series bar chart and the third and fourth views are bar charts.
- **Interactive Idiom(Mutli-form, overview & detail):**

(1) The first view is linked bidirectionally to the second and third view. By selecting locations on the map, the user can see the gender, ethnicity distribution as well as the top 5 objects of search. Similarly, the user can select values for the gender, ethnicity and object of search and filter the values being showcased on the map. The user can filter on age via an interactive checkbox as well.

(2) The second and third view are also linked bidirectionally. The user can perform a set of hierarchical filtering operations i.e select a particular value for the location attribute and then filter on gender/ethnicity to see how it affects

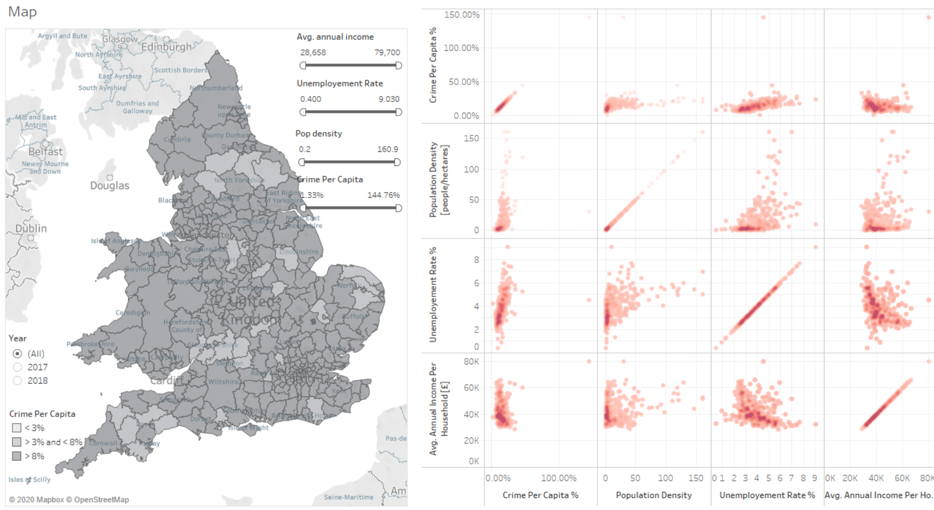

Figure 8: Interaction of features across the country

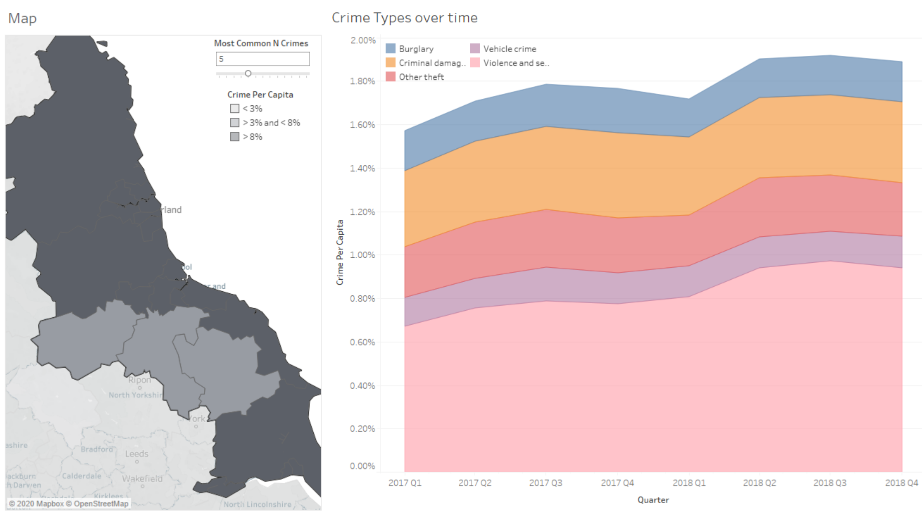

Figure 9: Evolution of attributes over time for the Norther coastal areas

the distribution of the top five objects of search and vice versa.

(3) Finally, the third view is linked unidirectional to the fourth view. The user can select values from the object of search attribute and see the top seven outcome distribution.

A point to be noted is that the third and fourth views are linked in an overview-detail manner as they share the same visual encoding.

4.4 Stop & Search - Task 2

This task corresponds to figure 5.

- **Goal:** The goal is to do an *exploratory analysis* for the top 5 overall objects of search as a function of time.
- **Means:** A search is performed to retrieve the values of each quarter (i.e *navigation*). The data is also filtered interactively by the user based on location, age, gender and ethnicity (i.e *re-organisation*).
- **Characteristics:** The *high level characteristics* are comparing trends through time.
- **Target:** The task is carried out on *temporal* and *spatial* relations via the location and date attributes linked to the search per capita corresponding to the top 5 overall objects of search. The user can also filter the data using age, gender and ethnicity.
- **Cardinality:** The task considers *multiple instances* for all the attributes involved in order to put the data into context with respect to time.
- **How is it shown:** The first view is a choropleth map and the second view comprises of line charts. The two views are

linked in a unidirectional manner such that it is of the form of a multiform overview & detail interactive idiom.

4.5 Stop & Search - Task 3

This task corresponds to figure 6.

- **Goal:** The goal is to perform an *exploratory* analysis on the objects of search for different locations.
- **Means:** The goal is accomplished via the use of *navigation* on the search per capita for each object of search and is *reorganized* based on interactive filtering via the selection of year, location, age, gender/ethnicity and finally *relation* is used to be able to compare for the different regions.
- **Characteristics:** The task's *high-level data characteristics* allow for comparison of distributions as well as outliers.
- **Target:** The task is carried out on *spatial relations* via the location attribute linked to the search per capita for each object of search. The user can filter based on age, gender,ethnicity and year too.
- **Cardinality:** The task investigates *all instances* for the object of search but the cardinality of the other attributes can be changed by the user.
- **How is it shown:** The first view in figure 6 is a choropleth map and the second is a multi-series bar chart. The two views are linked in a unidirectional manner such that it is of the form of a multiform overview & detail interactive idiom.

4.6 Crime - Task 1

This task corresponds to figure 7.

- **Goal:** The goal is to run an exploratory analysis via *undirected* search on crime type.
- **Means:** The goal is reached by means of *navigation* for the values pertaining to the different types of crime and by *re-organisation* of the data via a filter on locations on the map and the year.
- **Characteristics:**The *high level characteristics* include looking at distributions and the outliers across different areas in England.
- **Target:** The main target is the crime type attribute, this attribute is linked both spatially and temporally and can be filtered by the user via the location and year.
- **Cardinality:** We show the top seven types of crime(comparison) whereas the cardinality for location and year can be modified by the user.
- **How is it shown:** The first view in figure 7 is a choropleth map which is linked unidirectional to the second view, a bar chart via the use of a multi-form overview & detail interactive idiom.

4.7 Crime - Task 2

This task corresponds to figure 8.

- **Goal:** The goal of this task to allow the user to generate a hypothesis and is an *exploratory analysis*.
- **Means:** The goal is achieved via an exploration or *navigation* and filtering or *re-organisation* on many different attributes (annual income, unemployment rate, population density, crime per capita, year and location). The use of *relation* is also present in order to put the different attributes in context.
- **Characteristics:** This visualization highlights many different *high-level data characteristics* in terms of *dependency* and *correlation* between the above mentioned attributes as well as *frequency*.
- **Target:** The data relations for this task are spatial(data is linked to areas on the map), temporal(data is linked to time of year) and structural(attributes in the data are linked to each other) due to the many different filters that the user is able to personalize and the interplay between the associated attributes as well.
- **Cardinality:** The cardinality is chosen by the user and is initially set to show all instances of all attributes(overview).
- **How is it shown:** The first view in figure 8 is a choropleth map and the second view is a scatter plot matrix. Both views are linked bidirectionally to form a multiform overview and detail interactive idiom.

4.8 Crime - Task 3

This task corresponds to figure 9.

- **Goal:** The goal is to run an exploratory analysis on crime type considering the date attribute as well.
- **Means:** The analysis is made possible through *navigation* or a search on the crime per capita with respect to the date and filtering or *re-organisation* by selecting the locations on the map and the amount of most popular crime types.
- **Characteristics:** *high-level data characteristics* include the comparison of trends over time for different crime types and locations.
- **Target:** The data relation is *spatial* and *temporal* via the location and date attribute linked to crime per capita for the crime type attribute.
- **Cardinality:** The user can set the cardinality for location and crime type. The cardinality for the date is fixed to show all instances.
- **How is it shown:** The first view in figure 9 is a choropleth map and is linked unidirectional to the second view, a streamgraph such that it is a multi-form overview and detail interactive idiom.

4.9 Interactive idioms

(1) Hovering over any of the visual elements gives the user more detailed information via the use of the tooltip.
(2) Selecting one or more regions from the map acts as a filter for all views in the same visualization.
(3) There is also a parameter control in figure 3 and can be used to alter the top N most densely populated areas being shown.
(4) The radio buttons in figure 7 can be clicked to filter on year.
(5) The user can filter on age, gender and ethnicity by using the interactive drop-down lists in figures 4, 5 and 6.
(6) The slider in figure 9 can be used to alter the top N most common crime types being shown.
(7) The sliders in figure 8 can be used to filter on the shown attributes.

(8) Selection of different data points in the SPLOM in figure 8 highlights them on the map.
(9) The user can toggle the ordering of any of the multi-series or normal bar charts based on magnitude to on or off.
(10) The user can also change the viewpoint on the map using the pan and zoom via the interactive functionality provided by tableau.

5 THEORY BACKGROUND

The theory used to justify our choices mainly comes from the book and slides of the course. [4]

5.1 Box Plot

The box plot is used in figure 1. Generally, a box plot is a method for graphically depicting a quantitative attribute through their quartiles. In the case of our visualization, *box-and-whisker plot* would be a better way of defining it because of the lines extending horizontally that indicate the variability outside the upper and lower quartiles. In addition, the points that are outside of these boundaries are called *outliers* and have been considered with much interest by our group. [1] The line where the color goes from light grey to dark grey defines the median of the data, while the entire colored area contains data points between the first and third quartile. We also decided to show the average to demonstrate how it often differs significantly from the median. The more external boundaries are called *whiskers* and are drawn by measuring 1.5 times the interquartile range ($= IQR = Q3 - Q1$). A box plot is described using points as the marks and the channel used to describe the magnitude is position with respect to a common scale. Thus, the expressiveness principle and effectiveness principle are taken into account by using a highly ranked channel appropriate for a quantitative attribute.

5.2 Bar Chart(Separated, aligned and ordered)

The bar chart is used in figures 2, 4, 6 and 7.

Bar charts are typically used when representing a categorical attribute and a quantitative attribute along side it. For figure 2, we have manually defined ranges in the income values, which are known as bins and are the categorical attribute, whereas the frequency of points falling in those bins are used as quantitative attributes to give rise to a histogram.

Following the expressiveness principle, the categorical attribute is encoded using contiguous bounded areas in the form of lines distinct from each other and are separated horizontally. The quantitative attribute is expressed in terms of length and is used to order and align the lines with respect to the magnitude via a common axis. This also meets the effectiveness principle wherein we encode the most important attributes with the highly ranked channels. Bar charts are quite useful when performing look-ups and comparing values. We have also color coded the categorical attributes by using unique hues and provided a legend for ease of use while mapping the values of an attribute.

5.3 Multi-series bar chart(Separated, aligned and ordered)

A multi-series bar chart is used in figures 4, 6 and allows us to use two categorical attributes and one quantitative attribute that is associated with it. In our case, the categorical attributes are object of search and ethnicity in figure 4 and region in figure 6 whereas the quantitative attribute is search per capita. Similar to the bar charts, the categorical attributes are encoded using contiguous bounded lines that are distinct and separated horizontally but in the case of multi-series bar charts, there is an explicit hierarchy to the assignment of spatial regions such that there are two levels of separation, first at the level of ethnicity/region and then a second at the level of object of search. The quantitative attribute is encoded similar to the bar chart using the length of the lines aligned and ordered with respect to the magnitude employing a common axis thereby meeting the requirements of the expressiveness principle. And as always we strive to encode the most important attributes with the highly ranked channels to fit the effectiveness principles too. Multi-series bar charts can be better than stacked bar charts as a user can easily compare values which fall within a categorical attribute as the lines are all aligned on a common axis. Moreover, we also associate a color encoding to the different values of the categorical variables. As mentioned previously, we give a unique hue to each value and provide a color legend to make it more convenient for the user.

5.4 Geographical point map

The point map is used in figure 3. This visualization is used to represent a spatial attribute alongside two quantitative attributes. The marks used in the visualisation comprise of points. The channel used to encode the location(categorical attribute) are the different spatial regions on the geographical map. The channel used to encode the average annual income per household(quantitative attribute) is encoded using a custom sequential red color map. The perceptual ordering of the magnitude is implied via the change in luminance. Similarly the channel used to encode the unemployement rate is the size of the points. Therefore, we have abide by the expressiveness and effectiveness principles by matching the appropriate channels with right attributes(i.e categorical/quantitative) and have attempted to use the channels that are ranked high. However, a drawback of using size and color together is that if the size of the points are too small, the color becomes less apparent. In addition, the human visual system naturally experiences a disk's size in terms of its area. And the area of a disk, unlike its diameter or circumference, is not proportional to its radius, but to the square of the radius. For this reason, the disks' radius is corresponding to the *square root* of the value. Because this might not be immediately intuitive to the user, a legend has been provided on the map in order to easily compare the data points with each other. [2]

5.5 Choropleth Map

The choropleth map is used in figure 4, 5, 6, 7, 8 and 9. The use of such a map allows us to show a geographic heat map wherein the user can intuitively identify locations of high/low values and easily find outliers on the map by means of differences in colour. We believe that a choropleth map enhances the spatial relationship of the location attribute with the respective quantitative attributes associated with it.

The color encoding that we have employed for the quantitative attribute is of the form of a custom sequential grey-scale color map. This is because we need to give it a perceptual ordering via a change in luminance. Therefore as per the expressiveness principle, through the use of the choropleth map, we have used different spatial locations on a map as the channel to encode the categorical location attribute and the colour to encode the magnitude of the quantitative attribute associated with it. We have also tried to encode the most important attributes i.e location followed by respective quantitative attributes with the highest ranked channels as well, hoping again to meet the requirements of the effectiveness principle.

5.6 Line chart

The line chart is used in figure 5. The line chart is typically used to showcase two quantitative attributes. In our case, the two quantitative attributes are actually the quarters between the year 2017 and 2018 and the search per capita. The marks in this form of visual encoding are points with line connecting them. The channels used are of the form of points that are aligned on an axis via the length based on the magnitude of the quantitative attribute search per capita. The points are of course also separated and ordered by the key attribute which is date in this case into horizontal regions. Therefore, we again respect both the expressiveness and effectiveness principle by using the appropriate channels for our attributes as well as encode highly ranked channels with the associated attributes as well. The choice of a line chart is used here as there is an implicit ordering through the use of a clear temporal relationship via the date attribute with the search per capita attribute. The line charts are therefore ideal to show changes over time. We have used individual axes to represent the different values in the object of search attribute and have also included polynomial trend lines to better highlight the trends. This makes a very clear and concise visualization.

We have also color coded the different values in the object of search. Since it is categorical, we have used unique hues and provided a color legend for the respective lines in our visualization.

5.7 SPLOM

The SPLOM (scatterplot matrix) is used in figure 8. A SPLOM is a scatterplot matrix which makes use of multiple small views where there is a pairwise combination of attributes row and column wise. Each row-column set functions as a separate view. Each view has two quantitative attributes using points marks with both horizontal and vertical spatial positioning. Color is used to indicate the spread of the points whereby higher saturation equates to higher volume of points at the (almost) same position. This type of plot is used to determine if there is a correlation, clusters and trends between a series of variables.

In our example we have three quantitative attributes (Avg. Annual Income, Population density, Unemployment rate). Since all these attributes are quantitative it means they are ordered type attributes. For these attributes we make use of different saturation levels of the color red to show its quantity accordingly (higher saturation equals higher quantity). It makes spotting trends easier. This is how the expressiveness principle is applied in this example.

Apart from the expressiveness principle being applied, the effectiveness principle is also applied. Since we have two quantitative attributes that use an aligned spatial position, we use the most optimal ranking for ordered attributes. To provide extra information we also included color saturation which is also part of the ranking that shows the effectiveness principle for our graph.

Looking at the Unemployment rate vs Avg. Annual Income we can identify such a trend in which the higher income equates to a lower unemployment rate; conversely lower income equates to a higher unemployment rate. Normally a matrix would not necessarily have to include the matrices on the diagonal however in this case it was included to show the spread of the points which can be helpful.

5.8 Stacked area graph

The Stacked area graph is used in figure 9. This type of chart encodes three attributes where multiple sub-areas are stacked vertically and function as mark. The channels have a vertical spatial position channel for the quantitative attribute (Crime Per Capita) and the horizontal spatial position channel for the ordered key attribute (Quarter). The third key is a categorical key attribute (Crime types) to identify the different categories vertically.

When applying the expressiveness principle to these attributes the visualization shows that for the ordered key attribute we made use of an aligned spatial position; for the categorical key attribute we made use of color hue to define the different sub-areas into categories of its own. Every category has a distinct color such that there can be no mixing of categories.

Observing that the optimal for ordered attributes is to have an aligned spatial position with effectively an area and for categorical attributes it is to use the hue of the color we can state that the effectiveness principle also applies in our example.

Showing the total of all possible stacked sub-areas besides the set of selected categories would have made sense in the visualization if the sheer size of the total would not have been much bigger than the set of selected categories. Therefor the vertical quantitative attribute shows the percentage of the whole.

REFERENCES

[1] 2020. *Box Plot.* https://en.wikipedia.org/wiki/Box_plot
[2] 2020. *Bubble chart.* https://en.m.wikipedia.org/wiki/Bubble_chart
[3] Office of National Statistics. 2019. *UK Open Data.* https://www.ons.gov.uk/
[4] A. K. Peters. 2014. *Visualization Analysis & Design.* CRC Press.
[5] UK Police. 2019. *UK Crime and Policing Open Data.* https://data.police.uk/
[6] Hans-Jörg Schulz, Thomas Nocke, Magnus Heitzler, and Heidrun Schumann. 2013. A design space of visualization tasks. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2366–2375.