This document describes the preliminaries for past projects within the course IT740A

Example project 1

Can money buy happiness?

This project aims to analyze demographic information of different countries to determine the factors contributing to a country's "happiness rank", as defined by the UN World Happiness Report. A dataset was created through joining four separate datasets: a cost of living dataset compiled on kaggle through data on numbeo.com, the UN world happiness report, the World Bank GDP dataset, and the UN statistical database for country statistics. Different visualizations were chosen to attempt to highlight the similarities and differences between countries: scatterplots, bar charts, histograms, maps (with key performance indicators), and pie charts. Most visualizations were designed to be interactive in some capacity, utilizing brushlinking with other visualizations. A user study was performed to determine the efficacy of the visual analytics framework and determined the project to be generally successful overall, with some suggestions for improvements.

1. What is the effect of money (GDP or otherwise) on a country's overall happiness?
2. What factors contribute the most to a country's happiness?
3. What factors contribute most to your home country's happiness?
4. How does your home country differ from others?
5. What suggestions can you offer to improve your home country's happiness?

Data

Four datasets are analyzed:

1. Cost of living (in major cities, worldwide). Compiled on kaggle[1] with data submitted from numbeo.com, a collection of user-submitted statistics on cost of living.
2. World Happiness report. Compiled on kaggle[2] with data from the Gallup World Poll. This dataset contains the value for happiness index worldwide along with how much various other attributes contribute to a country's happiness index value. This is the same data that the UN published every year with their annual "World Happiness Report".
3. GDP per capita worldwide. Data compiled from World Bank[3].
4. Country Statistics. Compiled on kaggle[4] with data from the UN statistical databases[5]. This dataset contains various demographic information about each country, such as unemployment rate, estimated carbon emissions, and education levels.

[1] https://www.kaggle.com/andytran11996/cost-of-living?fbclid=IwAR2Mir-AFbCgsvk8Rlr3avedYXplY_Im1xx66PejxZafBHXHBwtEVCCU5_I#cost-of-living-2016.csv

[2] https://www.kaggle.com/unsdsn/world-happiness?fbclid=IwAR3Q2DcDXJCQM3S6XWnaHwkixmOuFr_EwgquB52kjLB0L9Nb5js3sQno4_o#2016.csv

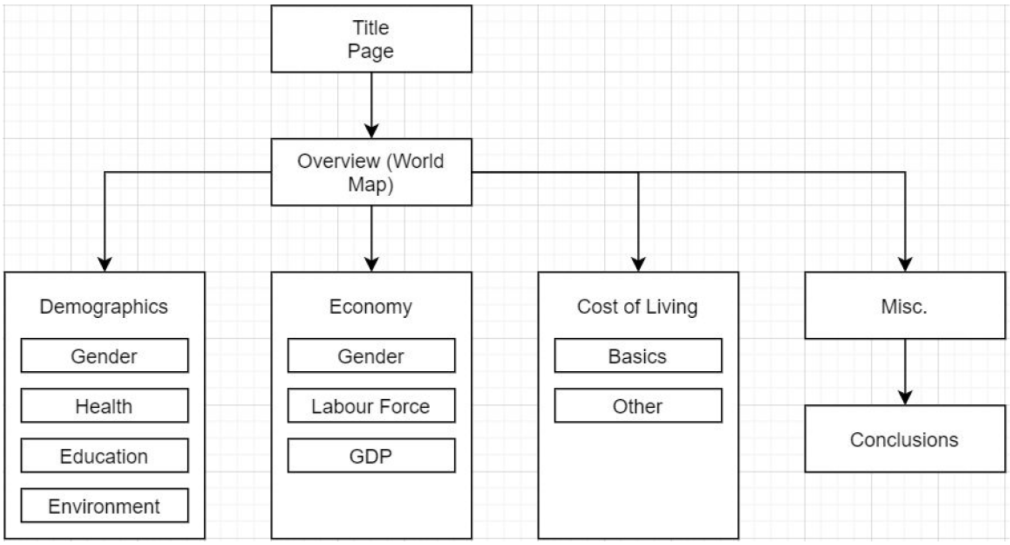
[3] https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

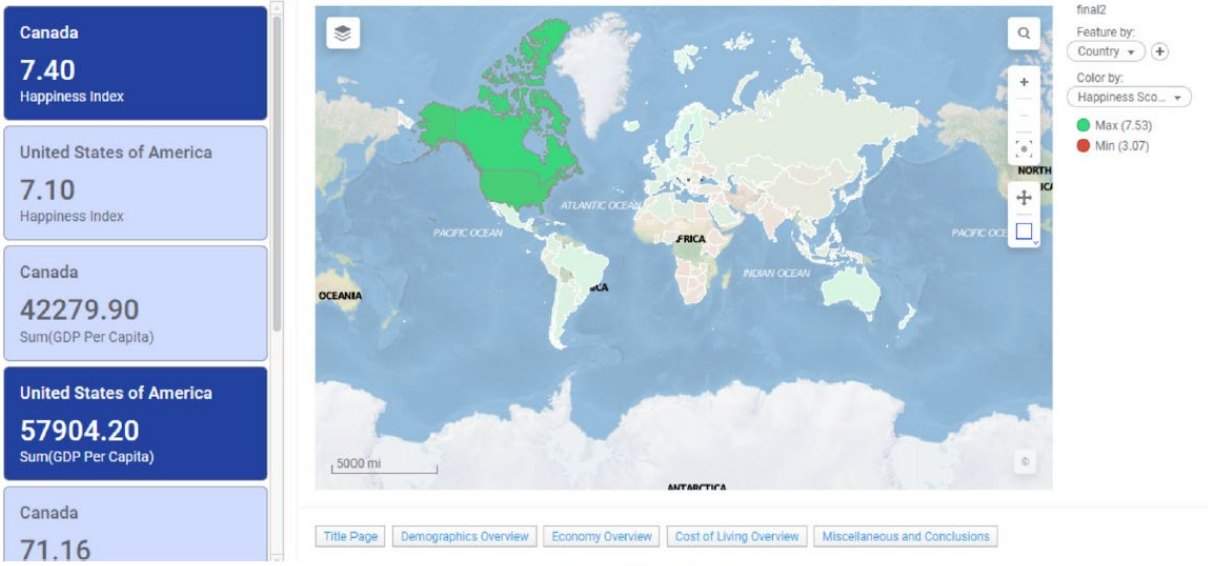
[4] https://www.kaggle.com/sudalairajkumar/undata-country-profiles?fbclid=IwAR0Q4FTeGwUjeGMGRatmnZabDESUNpP09sRG0IRA-ecDpZfew_ngI7ANECI

[5] http://data.un.org/

This document describes the preliminaries for past projects within the course IT740A

Example figures from the report




Canada
7.40
Happiness Index

United States of America
7.10
Happiness Index

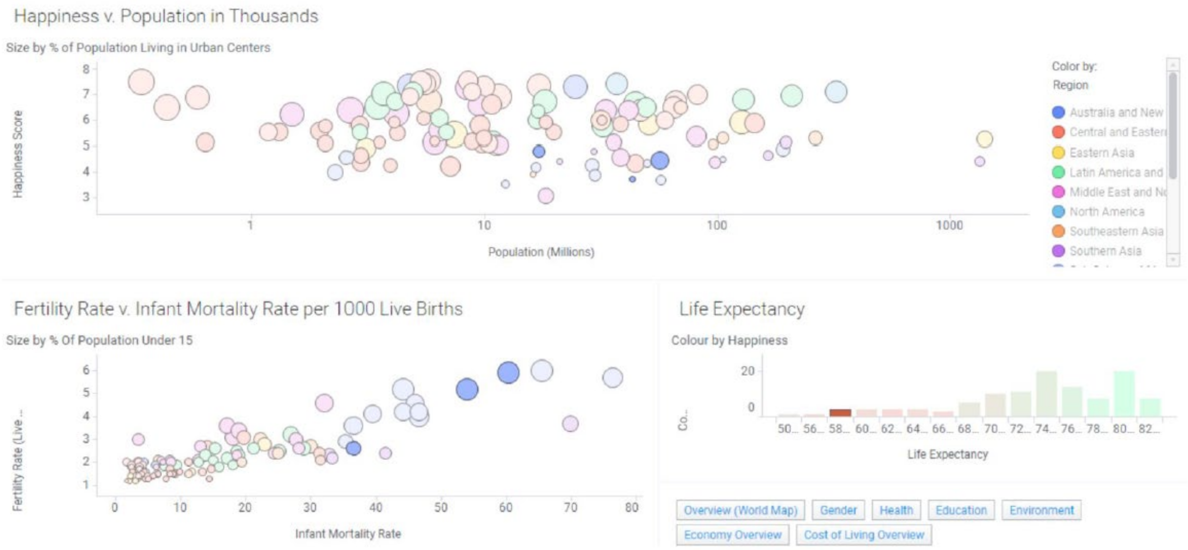
Canada
42279.90
Sum(GDP Per Capita)

United States of America
57904.20
Sum(GDP Per Capita)

Canada
71.16

final2
Feature by: Country
Color by: Happiness Sco...
● Max (7.53)
● Min (3.07)

Fig. 3: Overview (World Map)


Happiness v. Population in Thousands

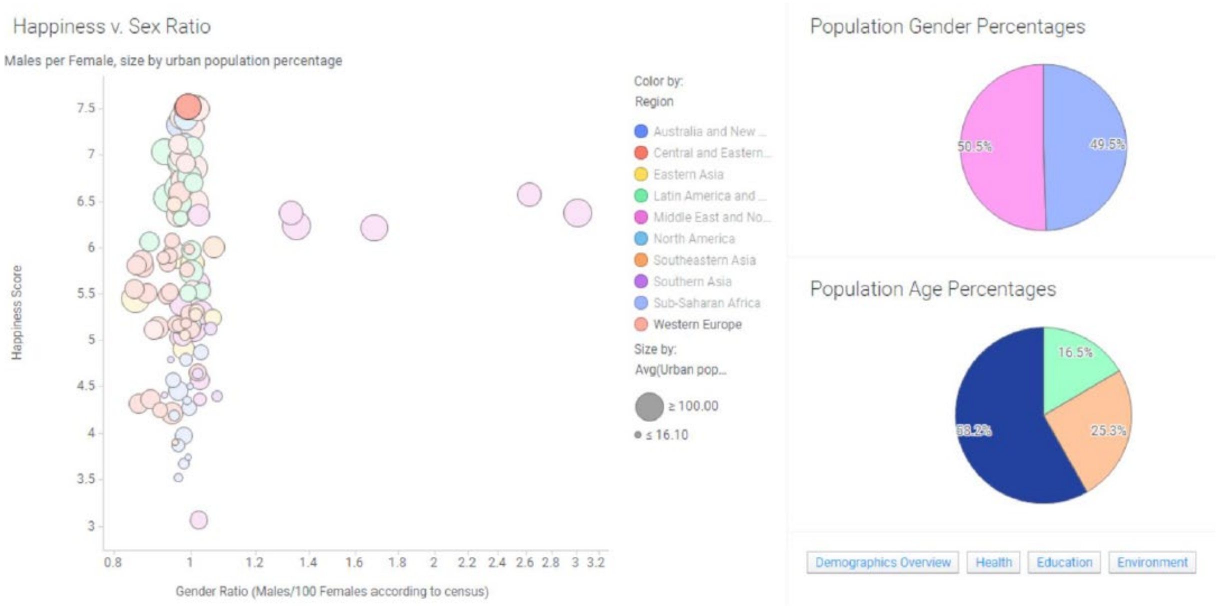
Size by % of Population Living in Urban Centers

Fertility Rate v. Infant Mortality Rate per 1000 Live Births

Size by % Of Population Under 15

Life Expectancy

Colour by Happiness



Example project 2

Economic and environmental aspects of solar energy

Over the past years, there have been many improvements following the efforts of sustainable energy usage – including people's awareness of using energy responsibly and an increase in green energy production. An important factor for this development is giving people a clear reason and incentive for executing a shift from their current energy provider to a green energy provider. As part of this, it is essential to understand current consumption and compare this to renewable energy potentials to decide if a change would be worth the investment. In particular, this project addresses the problem of making a decision of whether or not to switch from a conventional electricity provider to solar panels on the roof of one's own house. Thus, the following question should ultimately be answered:

- Will a switch to solar energy from solar roof panels make sense from an economical and environmental perspective?

In the process of doing so, the following questions should be answered as well:

- What is the current cost of electricity?
- What is the current electricity consumption?
- What is the solar radiation for a particular area in the USA?
- Given the solar radiation, what is the solar electricity potential?
- What costs are associated with the original investment of switching to solar panels?

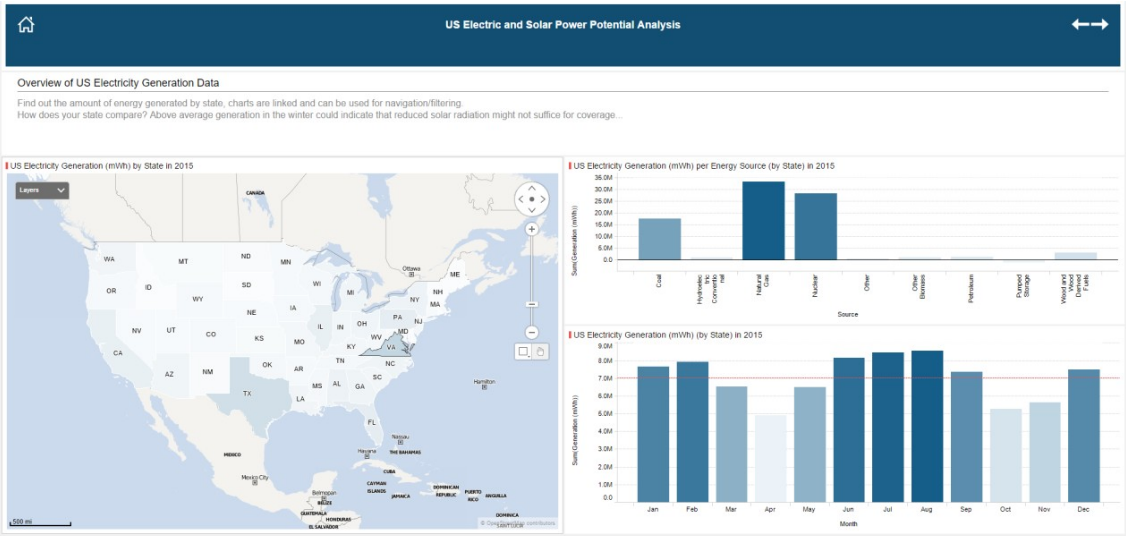
Data

Datasets to be used: US Electric Utility Rate Data (2015), two separate data sets including rate data from (non-) investor-owned companies by US zip code, to be merged to get complete pricing overview

This document describes the preliminaries for past projects within the course IT740A

1. US Electricity Generation Data (2015) by state per month, providing information about volume and source of energy generation[6]
2. US Electric Power Consumption Data, no detailed county/city/zip code data available, will use state-level data, averages per household/capita, possibly to be crawled from web[7]
3. Solar Radiation Data for USA, detailing solar radiation exposure on a city level, sources either National Solar Radiation Database (NSRDB)[8]
4. or files from EnergyPlus4 (depending on ability to handle file format)[9]

Example figures from the report



US Electric and Solar Power Potential Analysis

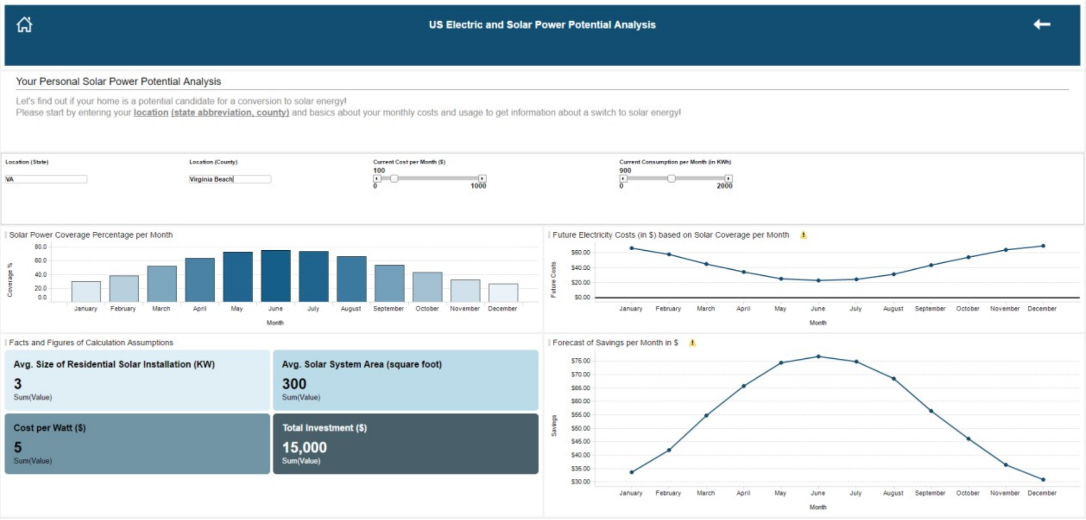
Overview of US Electricity Generation Data

Find out the amount of energy generated by state, charts are linked and can be used for navigation/filtering.
How does your state compare? Above average generation in the winter could indicate that reduced solar radiation might not suffice for coverage...

US Electricity Generation (mWh) by State in 2015

US Electricity Generation (mWh) per Energy Source (by State) in 2015

US Electricity Generation (mWh) (by State) in 2015



US Electric and Solar Power Potential Analysis

Your Personal Solar Power Potential Analysis

Let's find out if your home is a potential candidate for a conversion to solar energy!
Please start by entering your **location (state abbreviation, county)** and basics about your monthly costs and usage to get information about a switch to solar energy!

Location (State): VA

Location (County): Virginia Beach

Current Cost per Month ($): 100 (0 – 1000)

Current Consumption per Month (in KWh): 900 (0 – 2000)

Solar Power Coverage Percentage per Month

Future Electricity Costs (in $) based on Solar Coverage per Month

Facts and Figures of Calculation Assumptions

Avg. Size of Residential Solar Installation (KW)
3
Sum(Value)

Avg. Solar System Area (square foot)
300
Sum(Value)

Cost per Watt ($)
5
Sum(Value)

Total Investment ($)
15,000
Sum(Value)

Forecast of Savings per Month in $

[6] https://openei.org/datasets/dataset/u-s-electric-utility-companies-and-rates-look-up-by-zipcode-2015

[7] https://www.eia.gov/electricity/data/state/

[8] https://nsrdb.nrel.gov/

[9] https://energyplus.net/weather-region/north_and_central_america_wmo_region_4/USA%20%20

Example project 3

Soccer stats

Understanding the sport of soccer from an empirical perspective is challenging. Competitive matches are fast and dynamic, and contain a large variety of repeated, discrete events such as tackles, runs, passes, shots and goals. Any ensuing data therefore contains much freedom and randomness, thus making every match tricky to analyse. This project aims to apply the methodology of visual data analysis and incorporate data visualisation to better understand soccer from an empirical perspective, so that more robust performance analyses can be carried out. Specifically, this project seeks to better understand the different playing styles that are used by different teams, focusing on the following questions in particular:

- Do soccer teams always play the same way?
- Do soccer teams change how they play if they are playing at home versus away?
- Is there a relationship between the way a team plays and their final league position?
- How are different playing styles embodied in matches?

For clarity, consider the following. In soccer, the objective of each team is to score as many goals as possible, while conceding as few as possible. Each team has a maximum of eleven players on the field who work together to try and win the match, however teams can position their players in whichever way they please and operate according to whichever system best suits their strengths - so long as the team plays within the rules and laws of the sport. This 'system' may be thought of as a tactical, predetermined match strategy, or alternatively, as a reactive strategy that arises as the team attempts to cope with some superior opposition. Either way, the system, or 'playing style' is an emergent characteristic of the team, derived from the collective actions of the individual players. Modelling and interactively visualising these styles provides the opportunity for analysis, and hopefully, a more comprehensive understanding of how playing styles are influenced and how they can change.

A secondary aim of this project is to ensure that the visualisation produced can be used interactively by non-technical users (particularly soccer coaches) to further explore the data along similar lines of enquiry. Soccer coaches generally rely on live observation to inform their understanding of performance, yet due to the speed at which high-level soccer is played there is significant room for error. Coaches spend years developing their observational skills to reduce the impact of this, but even then, their judgement remains inhibited by the subjectivity of human observation. Effectively modelling and visualising performance data would provide coaches with an external tool to guide and inform their thinking, obtain greater confidence in their conclusions and ultimately be much better supported in their decision making.

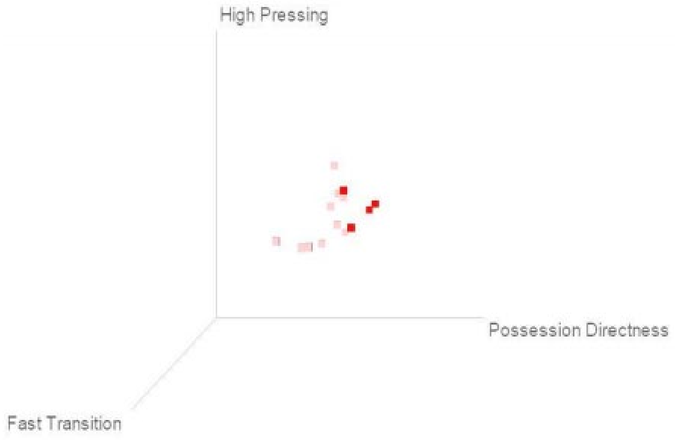
Data

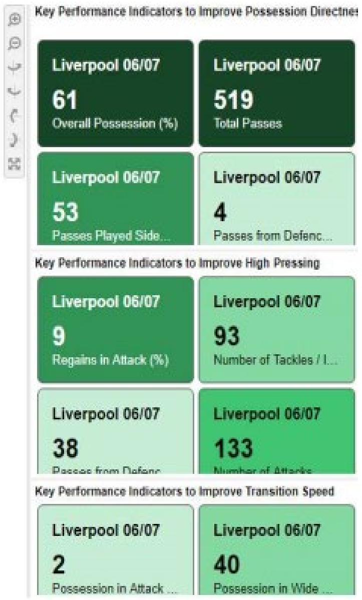
The dataset contains 196 records featuring match event data. The data was recorded during matches which took place during the English Premier League and the Spanish La Liga competitions of 2006/2007 and 2010/2011 using a player tracking camera system. Note that in major European soccer leagues, it is common for league competitions to begin in the autumn and continue to run until the following summer. The dataset was obtained from The Football Exchange, and is available on request.

This document describes the preliminaries for past projects within the course IT740A

[Teacher note: there are free sports datasets for ML here: https://lionbridge.ai/datasets/20-free-sports-datasets-for-machine-learning/)]

Example figure from the report





Key Performance Indicators to Improve Possession Directnes

Liverpool 06/07
61
Overall Possession (%)

Liverpool 06/07
519
Total Passes

Liverpool 06/07
53
Passes Played Side...

Liverpool 06/07
4
Passes from Defenc...

Key Performance Indicators to Improve High Pressing

Liverpool 06/07
9
Regains in Attack (%)

Liverpool 06/07
93
Number of Tackles / I...

Liverpool 06/07
38
Passes from Defenc

Liverpool 06/07
133
Number of Attacks

Key Performance Indicators to Improve Transition Speed

Liverpool 06/07
2
Possession in Attack ...

Liverpool 06/07
40
Possession in Wide ...

Example project 4

Swedish injuries

Visual analytics (VA) can contribute a lot to the areas of personal health, clinical healthcare and public health policymaking (Shneiderman, Plaisant & Hesse, 2013). The focus in this project will be on helping policy making. By analyzing data about injuries in Sweden during the last couple of years policymakers could identify trends and patterns that could show what actions need to be taken to help the population (and save the government money). VA tools combine analytics and interactive visualizations to support user's reasoning which makes them advantageous to several fields, one of them being public health (PH). Often analysis made of public health data needs to be shown to a diverse staff and VA tools can give the flexibility needed to present data to people with different backgrounds (Ola & Sedig, 2014). Ola and Sedig (2014) present four different ways that VA tools can be useful within PH:

1) Interactive visual representations: Gives the users opportunity to choose the most appropriate visual form for the task at hand.
2) Interaction: Users can control their dialog with information.
3) Automatically generate tailored reports for different groups of stakeholders.
4) Adjust tasks for different users (novice to learned)

The chosen dataset contains data about medical incidents per region, age-group and gender in Sweden 2001-2016. The medical incidents are injuries and poisoning and does not include illnesses. It would be interesting to see if there is a correlation between region, age group and gender. Can one see that a certain type of injury is more common in a certain age-group or in a certain region? This could give insight to if there is a problem in a specific region, ages-group etc. This is a large data set and using visual analytics could give insights by showing patterns in the data. The aim of the project is to show an overall view of the data and giving the user opportunities to interact with it. The focus will be on presenting the data in such a way that a novice could understand and answer the questions. In a larger project one might want to use different dashboards adapted to different users to answer more complicated questions. This project, however, will only focus on two or three questions and the goal is to make one or two dashboard(s) for each question. More specifically, this project will try to answer the following questions:

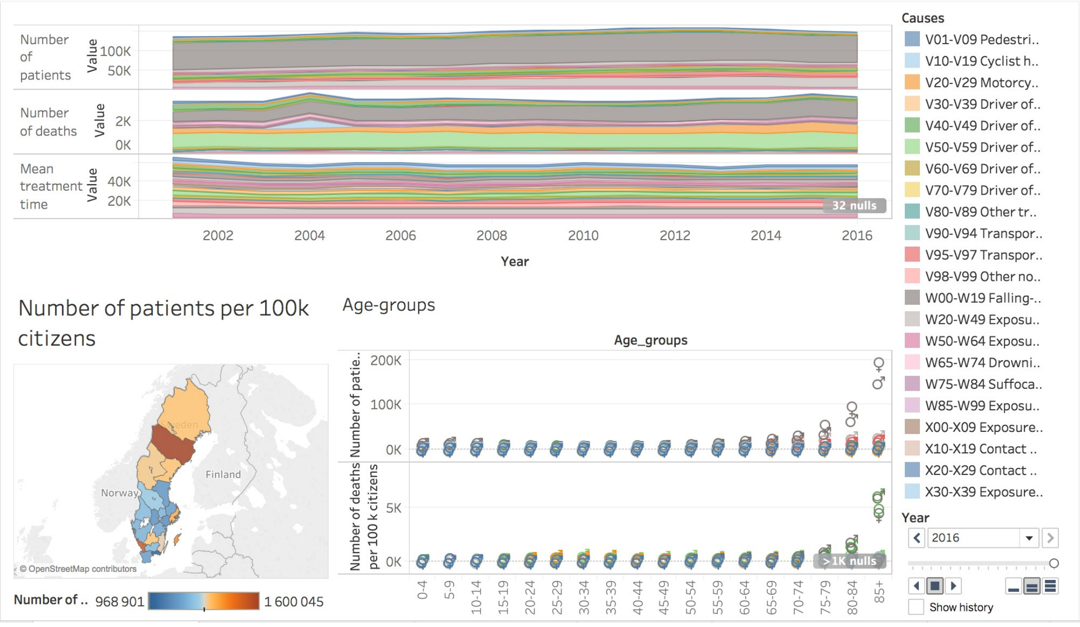
- Which types of accidents should we work on preventing to reduce the number of accidents over all? Are there any particular groups that we should focus more on?
- If we want more people to survive in Sweden which injuries should we work on preventing, in which age-groups and for which gender?
- Can a trend be identified?

Data

The data comes from the Swedish national board of health and welfare (Socialstyrelsen) (http://www.socialstyrelsen.se/statistik/forutvecklare) and the dataset contains several different measurements like days in hospital, number of times in hospital, number of people etc between the years of 2001-2016. This dataset was also combined with another data set from Socialstyrelsen to add two more variables (number of deaths and number of deaths per 100 000 citizens).

This type of data can be difficult to analyze because it is often geospatial and temporal (Shneiderman, Plaisant & Hesse, 2013). Luckily the data from the Swedish national board of health and welfare already has a pretty good structure. It is aggregated both on total, groups and per cause so the data shouldn't be too difficult to work with. A decision was made to combine two different datasets from Swedish national board of health and welfare and since these dataset are structured in the same way they are easy to combine.

Example figure from the report


Number of patients
Value
100K
50K
Number of deaths
Value
2K
0K
Mean treatment time
Value
40K
20K
32 nulls
2002
2004
2006
2008
2010
2012
2014
2016
Year
Number of patients per 100k citizens
Norway
Finland
© OpenStreetMap contributors
Number of .. 968 901
1 600 045
Age-groups
Age_groups
Number of patie..
200K
100K
0K
Number of deaths per 100 k citizens
5K
0K
>1K nulls
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
85+
Causes
V01-V09 Pedestri..
V10-V19 Cyclist h..
V20-V29 Motorcy..
V30-V39 Driver of..
V40-V49 Driver of..
V50-V59 Driver of..
V60-V69 Driver of..
V70-V79 Driver of..
V80-V89 Other tr..
V90-V94 Transpor..
V95-V97 Transpor..
V98-V99 Other no..
W00-W19 Falling-..
W20-W49 Exposu..
W50-W64 Exposu..
W65-W74 Drowni..
W75-W84 Suffoca..
W85-W99 Exposu..
X00-X09 Exposure..
X10-X19 Contact ..
X20-X29 Contact ..
X30-X39 Exposure..
Year
2016
Show history