Investigating the impact of automation on the personal, local and national economy

A Databases and Advanced Data Techniques project

Contents

Introduction

This investigation is led by my interest in joining the automation industry. A field where the principles of Computer Science are well applied.

As we are all well aware, automation on a large scale will lead to layoff jobs. These people are then left without an income, in turn impacting the local economy. This report aims to shed light into what an impact automation can have to the local economies and what amounts of money and GDP are affected.

After a good search, all I could find on open automation data were the probabilities of automation for each SOC job in the USA. The dataset also had the jobs per state for all SOC jobs. Using the information for all states I found GDPs and revenue values for those states and industries. Using the states as a common denominator.

In addition to that, a restriction made its self evident. The automation probabilities were calculated for the circumstances of 2012. Thus, all other datasets had to be for 2012 as well.

Background

Before we start, a few acronyms will be extensively used.

These are:
NAICS / SIC: short for North American Industry Classification System. America sorted all industries into a few types. These then have Sectors (the man types like Mining) and I called the details: sub-groups (e.g. Gas extraction and ore extraction). This is a very nice standard to classify all industries and businesses [5].

SOC: short for: Standard Occupation Classification. A system that sorts all jobs in America into a few main sectors (Like “Production Occupations”) and some sub-groups (like “Metal Workers and Plastic Workers” or “Plant and System Operators”). Simmilar to NAICS, just with job types instead of industries.

Stage 1

Data review

Automation data [1]:
The basis of this dataset was created in a paper investigating the future of employment (2013, C. Benedikt, et al). In their report they used machine learning that considered several job dependent factors. And it computed the probability of automation for 702 SOC jobs in the USA. Mr. Needs then added the number of jobs per state and published the dataset on data.world. In terms of quality: the probability calculation has been critiqued to be a bit simple. However, the mercies are sufficiently accurate for this exercise. The detail sufficient, highlighting the probability and number of jobs per state. Although good documentation is lacking. For a new comer it might be a bit confusing at first. When considering the openness and source of the data. It is definitely real, compiled with the genuine and unbiased intend to shed light into the future of job security. This resulted in it being used with several other studies as well. This also testifies to the availability, several other interested people managed to find it easily.

Economic census [2]:
This dataset was found on data.world but originated from the United States Census Bureau. This bureau serves to inform researchers about economic statistics. If this data should be biased or manipulated, the parties that intend to use it for economic analysis will have false economic reports. And if America does one thing right, it is to have a very healthy economy, this includes economic reports for analysis. So, it is safe to presume that this data is trust worthy. And since this data comes from and official bureau whose aim is to inform all parties interested, it was very easy to find the required data.

Both datasets from the Bureau of Economic Analysis [3] [4]:
As the source's name implies. These datasets originate from an official USA bureau. Resulting in all data being open, real and the intend of this data is to make informed decisions regarding the economy of the USA. If any bias and manipulated information are present in these datasets, economic analyses will be wrong and that will have severe consequences. Besides that, I used the datasets that are published a few years after 2012. At those stages the information was revised several times. They also describe the use as being fee for all. The aim of this data is to inform all that are interested. So, it was very easy to find by means of a quick online search for specific GDP values.

After discovering the economic datasets, the most important facts were cross-referenced with other datasets from the respective pages and all concluded to the same information. Some variations were detected, but that might the result of gathering data in different months on 2012. The economy never stands still.

The census data unfortunately works with the 2007 NAICS codes, but the differences to the 2012 revision are so small it really does not matter.

Minor complications were detected in the datasets from the Bureau of Economic Analysis. For example: some entries had random blanks, wrong data type and few two NAICS codes to one description mapping.

At the end, no data is perfect and some cleaning has to happen.

Questions

With the theme of local economic impact of automation. The following questions were conceived to gather some insight. These questions aim to use the automation dataset in combination with the economic datasets to try and determine the monetary and GDP impact.

1. What are the number of jobs lost with automation for each state?
2. What is the local economic impact of jobs lost due to automation?
   Impact in terms of revenue that is no longer available to the local economy. Considering that workers get a salary that was spend at the supermarket and other establishments. This revenue stream will now be interrupted. What are the consequences of this. Note, this does not mean that revenue disappears, it will just stay in the companies accounts or be spent elsewhere.
3. What are the differences between the GDPs affected in all industries and states?
   This aims to determine the highest affected industries and states.

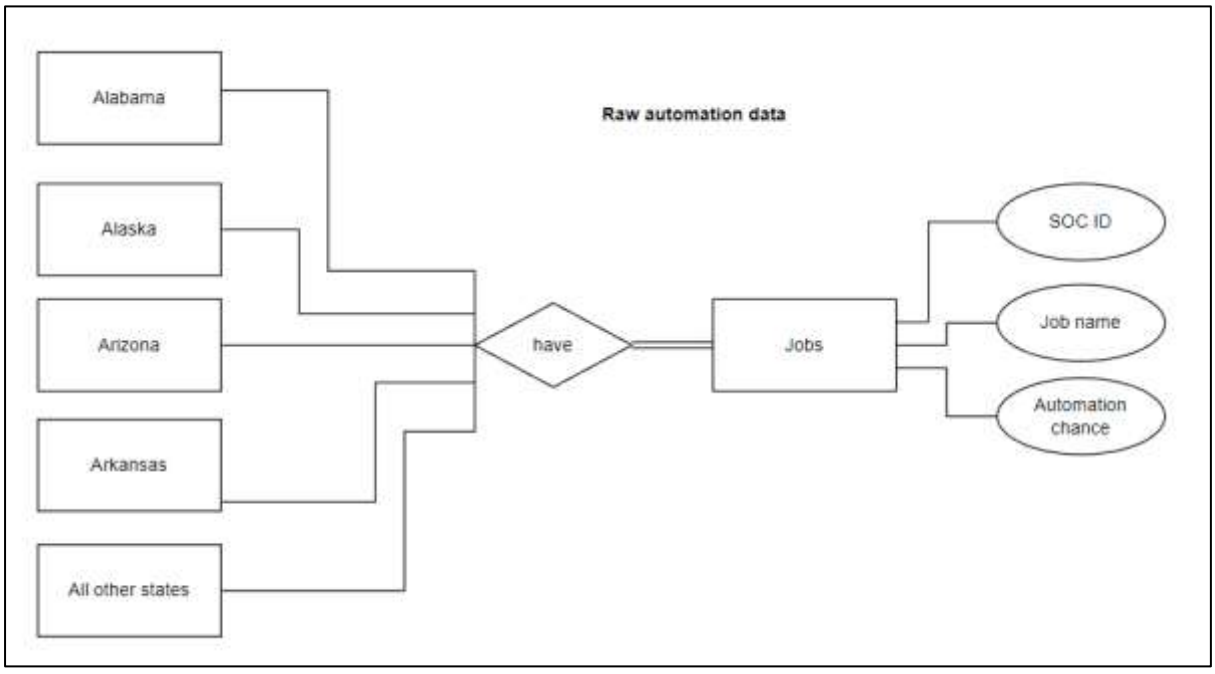
Stage 2

Entity Relationship Modeling

These are the ER-models for the 4 main datasets

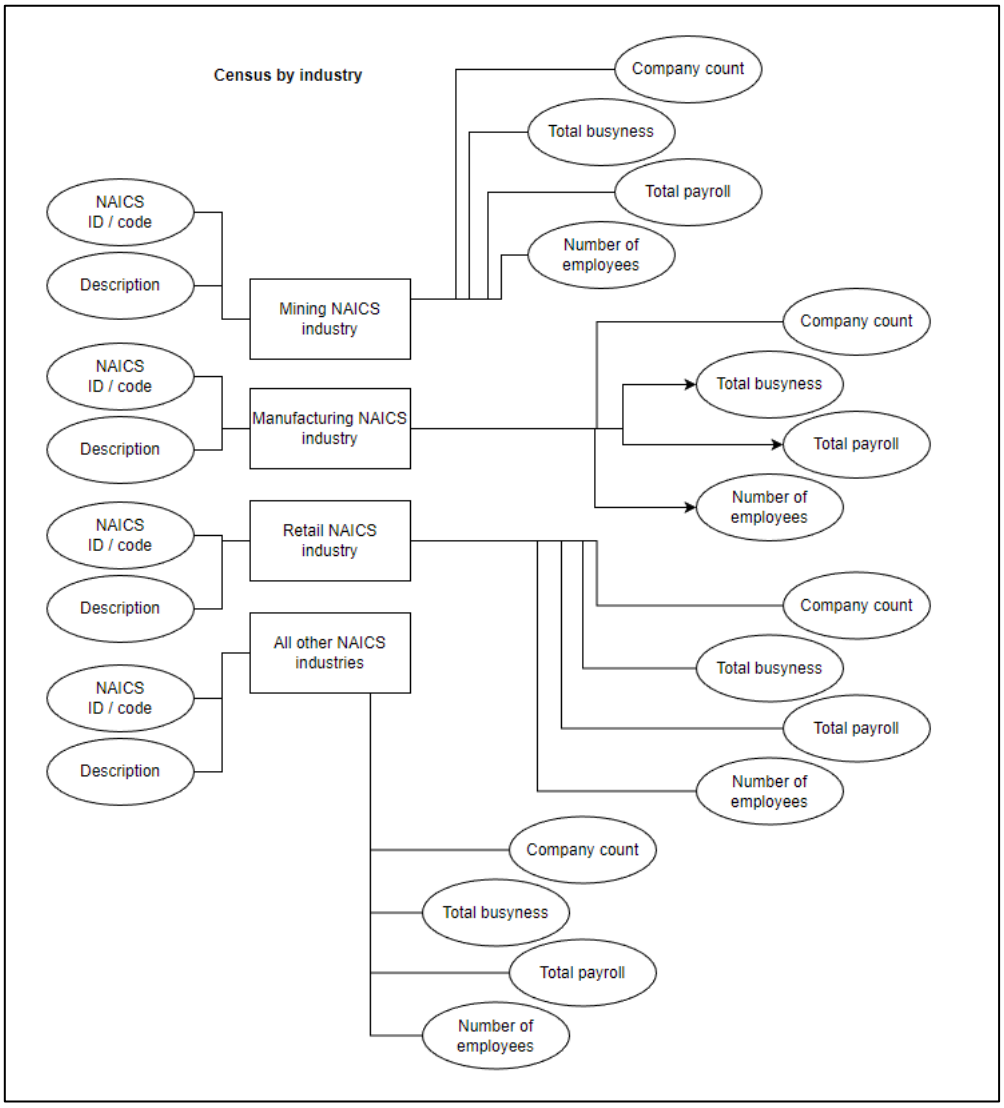
The automation dataset [1]

Please note: all states are regarded as their own entities but not all are listed.



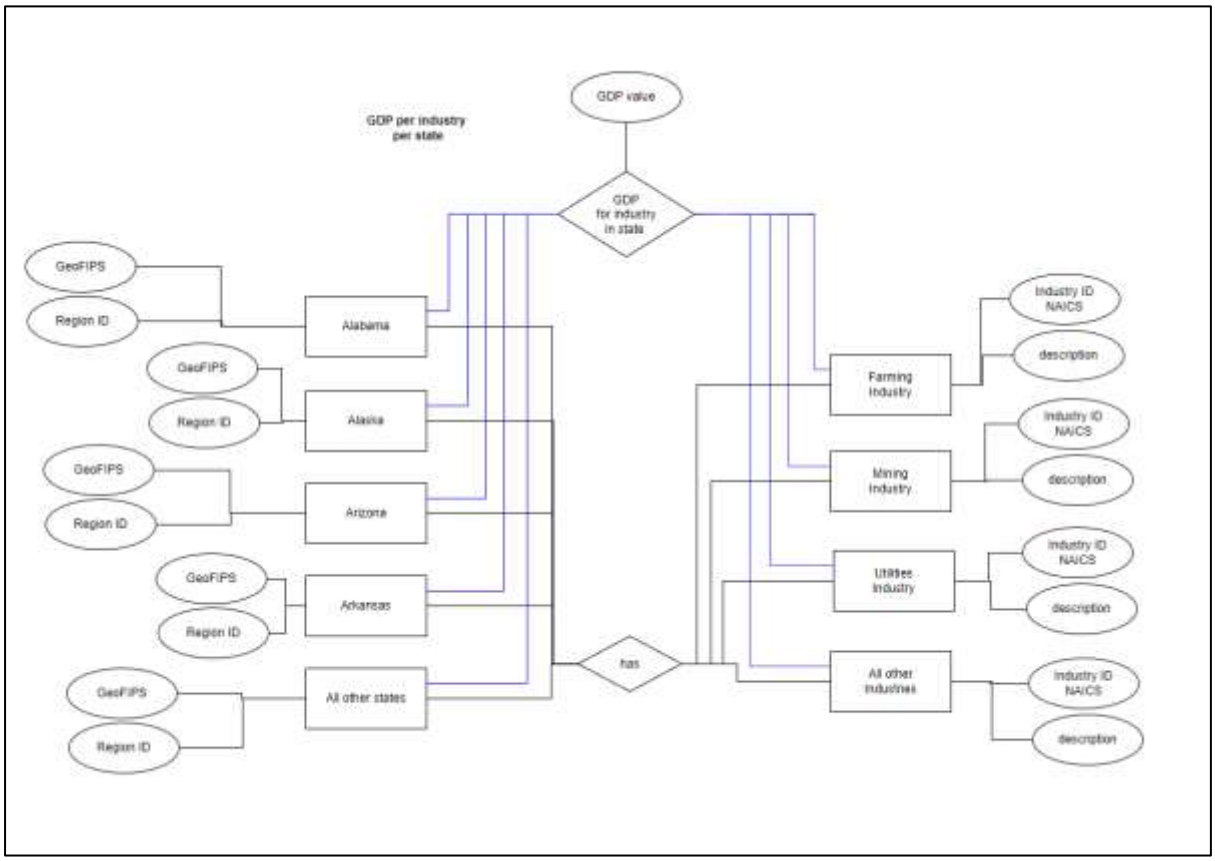
Economic census [2]

Please note: The central repeating entities represent all industries of the NAICS classification. I just depicted a few.



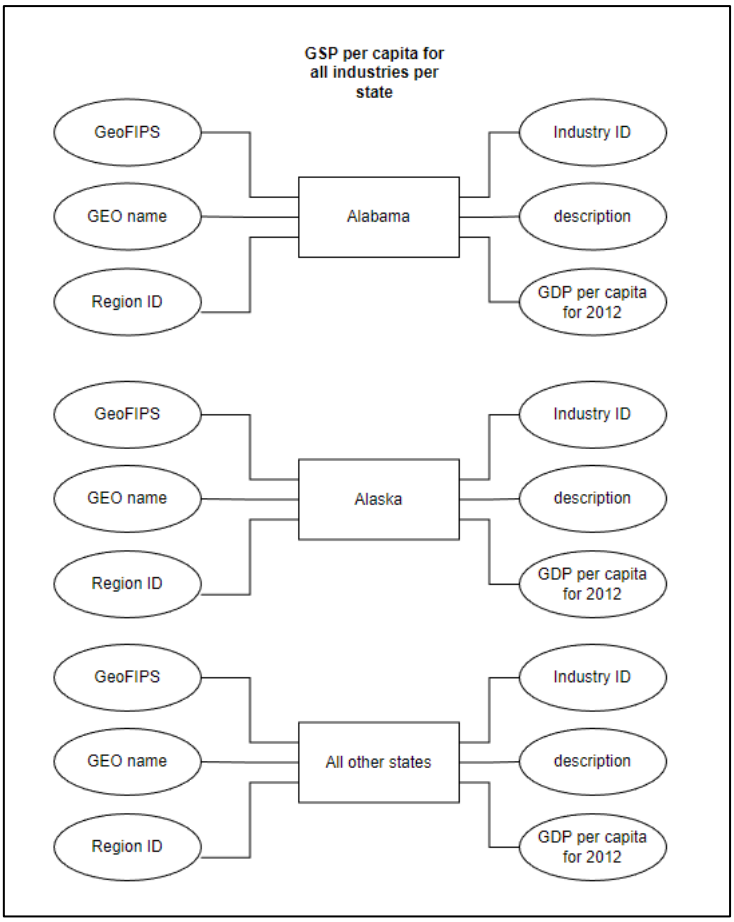
GDP per industry for all states [3]

One again I only depicted a few samples of all states and industries
And this dataset is a bit more complicated, thus I have indicated the 2 different relations with different colors.



GDP per capita for all states [4]

Here a few sample states were used to highlight the repeating entities.



Data cleaning & modification

In general, this included removed label fields, other year fields and after normalization, the redundant fields. Like state names, various IDs and descriptions.

Some datasets have a lot of information about years that are not 2012. Removing all of them allowed changes to the dataset and structure that were previously impossible.

To help understand the NAICS grouping and classification I used the official guide [5]. This was fundamental to the mapping of SOC jobs to NAICS industries and the general understanding of the datasets

Below are the details for cleaning, during cleaning 1NF was already considered:

Raw automation data

- Nothing to change

Economic census

- Remove range industry ids (some ids had a starting and ending id), we don't need that much detail
- Remove all commas in the number fields

- These steps did cause a bit of information to be lost, but not enough to cause serious misinformation.

GPD per industry per state

- Remove all years, just keep 2012
- Remove Component id
- Removed all entries that are not related to GDP by state. They include other economic measures, not relevant to this investigation
- GDP values have letters or no value, replaced with 0
- Industry classification filed: removed the end of some ranged fields. We don't need that much detail

GDP per capita for all states

- Removed all fields not relevant

Normalization

1st normal form

Raw automation data

- All fields have atomic data, from the start
- Primary key: SOC number, the value per state depend on the SOC type

Census

- Data is atomic, after a bit of cleaning
- Primary key: NAICS code, all fields need the NAICS code to identify in what industry how many businesses are, for example.

GPD per industry per state

- Description is a text field and can not be split apart without losing information
- Other than that, all data is atomic with no lists or the like.
- Primary keys: GeoFIPS & Industry ID, only those combined allow the unique identification of what industry in what state is being considered. State and industry alone do not work.

GDP per capita for all states

- All data is good.
- Primary key: GeoIFPS (the ID for all states). Using this field allows the unique identification of the GDPs for all states

2nd normal form

Raw automation data

- Job name can be removed into its own table, we only need the SOC code and the data for every state. Job names will be redundant data.
    - This means a new table is created: SOC
    - It contains SOC code, Sector code, subgroup code and description

Census per state and industry

- Remove the NAICS description. It is not needed in the context of numbers per industry.
    - This will create a new table: NAICS
    - It will contain the NAICS code and the description

GDP per capita for all states

- Geo name and region can be removed so we only have the id and GDP data. They are not important in relation to the GDP per capita.
    - The states table will be created
    - It will contain the State GeoIFPS, GeoName and the Region the state belongs to

GPD per industry per state

- Here we can remove the GeoName and Region (redundant data with the States table)
- And we can remove the industry classification and description. This will be redundant since we have a NAICS table.
- All fields removed are not needed in this context and do not depend on the primary key.

3rd normal form

No table has any transitive dependencies.

BCNF

No tables have anything to improve

4th normal form

There could be something to improve but that means taking the repeating state data and giving them their own fields. This is a horrible idea since then all states have to be individually addressed during the query process. Or, even worse, a separate query has to be made just to get all states. And they are not few.

New and junction tables:

This includes efforts to build them and mentions some additional error correcting in the datasets.

- SOC:
    - Fields: SOC code & description
- NAICS:
    - Fields: NAICS code, industry id, description
    - Adding farming sector 11
    - Adding “all industries” and “private industries”
    - Method: lookup in Excel, blanks in NAICS descriptions are added with new NAICS codes, we want the numbers with roughly the right industries. Labeled with “(new field)”
      A lot of manual work but also some with the lookup function or find and replace.
    - Inconsistencies between the 2007 and the 2012 NAICS codes could be the cause to the differences.
- States:

    - Fields: code (GeoIFPS), GeoName and region
- Industry to job mapping
    - This is the work of my own. During the search for a dataset, I thought that NAICS was the same as SOC (NAICS can sometimes be referred to as SIC). As it turned out, it is not so. This will, to a large extent, will be the foundation of the investigation. Solution is to build a junction table to correlate the SOC jobs to the NAICS industries. And this worked astoundingly well.
    - E.g.: Construction managers into the construction industry, Medical and health services managers into the health care industry.
      Some are a bit iffy. So, they go into the other services category.
    - This approach is not 100% accurate but I'm confident enough that the mapping is accurate to deliver correct insight.
    - Having said that, there are complications with some jobs. For example: secretaries are needed in all industries yet I placed them in one. The best would be to get someone with deep insight into these classifications and use their expertise. Better yet, have several mappings focusing on different aspects.

Stage 3

Database code

Field width specification has been omitted or left default where the width is not easily identifiable and where the data type will deprecate such feature soon.

```
CREATE DATABASE usa_automation;


CREATE TABLE SOC (
   code char(7) PRIMARY KEY,
   sector tinyint NOT NULL,
   subgroup smallint NOT NULL,
   description varchar(255) NOT NULL,
   probability float NOT NULL,
);

CREATE TABLE NAICS (
   code int PRIMARY KEY,
   industry_id mediumint DEFAULT NULL,
   description varchar(255) NOT NULL
);

CREATE TABLE states (
   code mediumint PRIMARY KEY,
   GeoName varchar(255) NOT NULL,
   Region tinyint NOT NULL
```

```
);

CREATE TABLE automation_data (
    SOC char(7) REFERENCES soc(code) ,
    states varchar(255) REFERENCES states(GeoName) ,
    jobs mediumint DEFAULT NULL,
    PRIMARY KEY (SOC, states)
);

SOC (jobs) to NAICS (industry) mapping, my data
CREATE TABLE soc_to_naics (
  soc_name varchar(255) PRIMARY KEY,
  sector tinyint DEFAULT NULL
);


Census data by industry
CREATE TABLE data_by_industry (
    naics_code int PRIMARY KEY REFERENCES NAICS(code),
    busyness_count int DEFAULT NULL,   -- number of businesses in set industry
    total_busyness bigint DEFAULT NULL,   -- total value of busyness (sales,
                                          -- shipments, receipts,
                                          --  revenue or other busyness)
    worker_count int DEFAULT NULL,
    annual_pay int DEFAULT NULL
);


GDP per state per industry
CREATE TABLE state_industry_gdps (
industry_id tinyint REFERENCES naics(industry_id) ,
       states varchar(255) references states(GeoName),
    gdp int DEFAULT NULL,
    PRIMARY KEY (industry_id, states)
);


GDP per capita for all states
CREATE TABLE gdp_ps_all_naics (
  geo_ifps mediumint PRIMARY KEY,
  total_gdp int NOT NULL
);
```

Populating the tables

Ideally table population would be the LOAD DATA command but it won't work well on my system and with laggy internet, working on the labs was less than productive. So, I used the import wizard from MySQL workbench and if that did not work, I used the CSV to SQL tool [6].

The load data command would have been something like:

```
LOAD DATA INFILE "../midterm/data/SOC codes processed"

        INTO TABLE SOC

        FIELDS TERMINATED BY ','
```

```
ENCLOSED BY '"'

LINES TERMINATED BY '\n'

IGNORE 1 ROWS; -- the headings
```

Querying the database to answer the questions

A better formatted view of the queries can be seen in the app.js file of the webapp.

The use of field aliases is for the webapp where one bar-graph helper function is used to process all input. So, the input has to be the same, but with different data and description.

Questions

1. What are the number of jobs lost with automation for each state?
    a. For this only the automation table will be considered. Multiplying the probability that a job will be lost, with the number of jobs per state.
    The Query:

```
SELECT automation_data.states as label, sum(round(automation_data.jobs*soc.probability)) as data
        FROM automation_data
        LEFT JOIN SOC
        ON automation_data.SOC = soc.code
        GROUP BY automation_data.states
        ORDER BY data DESC;
```

NOTE: 0 job losses will be due to no job info or less than 10 being available

2. What is the local economic impact of jobs lost due to automation?
    Here we consider the GDP per capita and the annual pay.
    a. For state GDP affected we use the previous query (total jobs lost) x GDP per capita in all states (and all industries)
    The full query

```
SELECT states_jobs.states as label, states_jobs.jobs_lost*gdp_ps_all_naics.total_gdp as data  -- (x)
   FROM (
      SELECT jobs_lost.states, jobs_lost.jobs_lost, states.code   -- getting the state codes
         FROM (   -- sub selects from https://stackoverflow.com/a/1888845
            SELECT automation_data.states,  -- getting all jobs lost and their states
sum(round(automation_data.jobs*soc.probability)) as jobs_lost
               FROM automation_data
               LEFT JOIN SOC
               ON automation_data.SOC = soc.code
               GROUP BY automation_data.states
               ) as jobs_lost
            JOIN states
            ON jobs_lost.states = states.GeoName
            ) as states_jobs
         JOIN gdp_ps_all_naics
         ON gdp_ps_all_naics.geo_ifps = states_jobs.code
         ORDER BY data DESC;
```

b. To get the total amounts of pay that are no longer available for each job type we use the total jobs lost query x annual pay from the economic census.
The query:

```
SELECT jobs_sector.description as label, data_by_industry.annual_pay*jobs_sector.jobs_lost as data
  FROM (
    SELECT soc_description.jobs_lost, soc_description.description, soc_to_naics.sector FROM (
      -- getting the naics sector from the SOC to naics mapping
      SELECT soc.code, jobs_affected.jobs_lost, soc.description FROM (  -- getting SOC description
        SELECT automation_data.soc,   -- getting total jobs lost
sum(round(automation_data.jobs*soc.probability)) as jobs_lost
          FROM automation_data
          LEFT JOIN SOC
          ON automation_data.SOC = soc.code
          GROUP BY automation_data.soc
          ) as jobs_affected
          JOIN soc
          ON jobs_affected.soc = soc.code
        ) as soc_description
        LEFT JOIN soc_to_naics
        ON soc_to_naics.soc_name = soc_description.description
        ) as jobs_sector
      JOIN data_by_industry
      ON data_by_industry.naics_code = jobs_sector.sector
      LIMIT 20
      ;
```

Note: this is over all jobs and nationwide, labeled by industry (small change for totals: sum and group by on industry but summing will lead to too large numbers)

3. What are the differences between the GDPs affected in all industries and states?
    a. Here we get the data from the GDP per industry and see how many jobs are lost per industry. Thus, getting the GDP affected by industry grouped by states.

```
SELECT jobs_n_industies.industry_name as label,
sum(jobs_n_industies.jobs_lost*state_industry_gdps.gdp) as data  FROM (  -- multiplying GDP and
jobs lost
  SELECT jobs_n_sectors.jobs_lost, jobs_n_sectors.description as job_name, jobs_n_sectors.states,
naics.industry_id, naics.description as industry_name FROM ( -- all SOCs for all states with labled to
what industry_id  (NO 1 & 2 id)
    SELECT jobs_affected.jobs_lost, jobs_affected.description, jobs_affected.states,
soc_to_naics.sector FROM (
      SELECT automation_data.soc, round(automation_data.jobs*soc.probability) as jobs_lost,
soc.description, automation_data.states
          FROM automation_data
          LEFT JOIN SOC
          ON automation_data.SOC = soc.code
        ) as jobs_affected
        JOIN soc_to_naics
```

```
    ON soc_to_naics.soc_name = jobs_affected.description
  ) as jobs_n_sectors
  JOIN naics
    ON naics.code = jobs_n_sectors.sector
  ORDER BY industry_id asc
) as jobs_n_industies
  JOIN state_industry_gdps
  ON state_industry_gdps.states = jobs_n_industies.states
     and state_industry_gdps.industry_id = jobs_n_industies.industry_id
GROUP BY jobs_n_industies.states, jobs_n_industies.industry_name
  having states = "Alabama"
  ORDER BY data DESC
;
```

DB reflection

The data is not perfect, as mentioned. I had to improvise with the new NAICS fields. Quite a number of fields had erroneous data from my understanding. This is economic data of the best performing economy in the world. I think they know that they are doing. Thus, I would have to investigate further to understand the data even better.

Second issue would be the SOC to NAICS mapping. It leaves a lot to be desired. My efforts were proper but during the sorting of 702 job types into the best fit industry. Some errors will have occurred.

Due to the time pressure, small imperfections could not be resolved during the data imports into the tables. The only way to resolve this would be to investigate substantial amounts of time to really clean the data. And thus, ensuring immaculate data can then serve even a professional purpose.

When it comes to the structure of the database. The way I implemented the tables and modified them from the original data serves to compile simpler queries. Considering that 59 geographical identities (states + regions) exist, this will be a feat to engineer the queries to use the more normalized structure. This is something I realized when writing the first select statements to answer the questions.

Despite all that, I feel confident that my efforts yielded factually correct insight into the effects of automation on the economy and that the research questions are answered in a way that makes them relevant and informative. Unfortunately, there is small degree of error preventing this investigation to be used as publishable insight. Yet, as stated, these minor errors can be easily corrected with some in depth cleaning.

Stage 4

This web application was developed on my local machine. Including the fully implemented MySQL database to serve the webapp. The webapp its self uses D3.js to visually present the data and the interactions allow sorting of the data.

The database

With the database build and populated. The mysqldump command was used. With that file the database on the labs was build. Thus, having identical databases.
There were a few complications but help from sqlshack [7] was very appreciated.

Web app

Fundamentally the node.js application is based on the demo provided by gregjopa on GitHub. The D3-server-side-demo repository served as the base on which the whole app was build. I added my own changes to customize for my needs. But, the versions of the libraries (I could not get any later versions to work without rewriting everything to make gregjopa's demo compatible with modern syntax) and way the app renders the pages comes from him.
Then the routs and MySQL logic was used from the provided sample app. I then just add the routes and customized to my needs.

Here is the link to the labs (The project is built in the "/home/coder/project" directory)
https://hub.labs.coursera.org/connect/sharedgtziwugy?forceRefresh=false&path=%2F%3Ffolder%3D%2Fhome%2Fcoder%2Fproject&isLabVersioning=true

I did some testing and the labs should work. But if anything should happen, please see the appended screenshots of the app working on my local machine.
If interested, the labs contain all code to build a database and run the code locally.

Conclusion

After the data sourcing, cleaning, modeling, structuring, querying and presenting. We can finally deduce the answers to the research questions

1. What are the number of jobs lost with automation for each state?
Listing the number of jobs for all states is futile. Thus, the state with the highest losses is California with 8,334,439 affected jobs. And the least hit state is Wyoming with 154,315 jobs affected.
In these numbers the state population plays a huge role.

2. What is the local economic impact of jobs lost due to automation?
    a. The total GDP from the people that is affected
    The highest loss for all people is in California with 440,350,084,565 and the lowest loss is in Vermont with only 6,797,039,634.
    Now these numbers represent the sum of GDP that the workers will no longer produce.

    b. The total amount of pay that is no longer available for each job type.
    Least losses have the Choreographers with a total of $896,722,582 annually and the Retail Salespersons loose the most with roughly $1,537,364,789,484,75,150, then followed by cashiers. These are the sums of all salaries that are affected. This means these amounts of money won't be spent in the national economy anymore.

3. What are the differences between the GDPs affected in all industries and states?
The retail trade loses the most with 3,138,976,030 and the least affected industry is mining, quarrying, oil and gas extraction with a GDP loss of 5,785,456.

This is all hypothetical estimation! Not all jobs can disappear overnight, automation is a gradual phasing out of people. And they will find other places to keep themselves productive. Additionally,

these measures were very “rough” at times. So not a guide, just some insight into the possible affects. This can be accurate work but only after the mentioned changed have been applied.

Concluding this project, one can say that it did achieve what it set out to accomplish. The Questions were answered and the insight was obtained. The Database is functional and has the ability to serve for purposes beyond this exercise. Extending and updating has become optimized as well.

Having said that, mistakes were made. The shape of the datasets played too much of an influence on the database design. And cleaning the data more thorough would have yielded more accurate results.

As next steps I would rework the input data, cleaning it to perfection. Then the information obtained from the queries will be 100% accurate. And after that the relational modeling will benefit from some adjustments.

References

Data sources

[1] Nedds, W. (2023) Occupations by State and Likelihood of Automation - dataset by wnedds. https://data.world/wnedds/occupations-by-state-and-likelihood-of-automation (Accessed: December 5, 2023).

[2] Economic Census by Industry - dataset by garyhoov (2023). https://data.world/garyhoov/economic-census-by-industry (Accessed: December 5, 2023).

[3] GDP by State and Industry, all components: US Department of Commerce, BEA, Bureau of Economic Analysis (2013) BEA : Gross domestic product by State, advance 2012 and revised 2009-2011. https://apps.bea.gov/regional/histdata/releases/0613gsp (Accessed: December 5, 2023).

[4] GDP by State and Industry, NAICS, Per Capita Real GDP: US Department of Commerce, BEA, Bureau of Economic Analysis (2013) BEA : Gross domestic product by State, advance 2012 and revised 2009-2011. https://apps.bea.gov/regional/histdata/releases/0613gsp (Accessed: December 5, 2023).

Code

Gregjopa (2020) GitHub - gregjopa/d3-server-side-demo: Render d3 visualizations server-side. https://github.com/gregjopa/d3-server-side-demo (Accessed: December 12, 2023).

Multiple group by: Is it possible to GROUP BY multiple columns using MySQL? (2009). https://stackoverflow.com/a/1841433 (Accessed: December 16, 2023).

sub selects from What is the error 'Every derived table must have its own alias' in MySQL? (2014). https://stackoverflow.com/a/1888845 (Accessed: December 13, 2023).

x label rotation from: Noob, D. (2013) How to rotate the text labels for the x Axis of a d3.js graph. http://www.d3noob.org/2013/01/how-to-rotate-text-labels-for-x-axis-of.html (Accessed: December 14, 2023).

HTML forms (no date). https://www.w3schools.com/html/html_forms.asp (Accessed: December 17, 2023).

Additional

[5] Understanding and grouping the NAICS industries: North American Industry Classification System (NAICS) U.S. Census Bureau (no date). https://www.census.gov/naics/?58967?yearbck=2012.

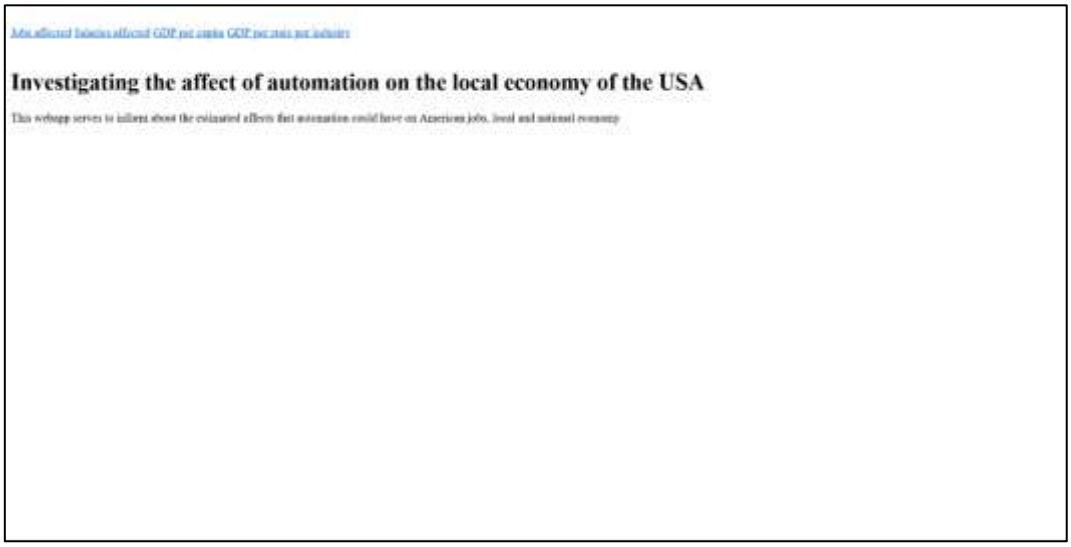
[6] CSV to SQL insert tool: CSV to SQL Converter (no date). https://www.convertcsv.com/csv-to-sql.htm (Accessed: December 8, 2023).

[7] SQL dump: Upadhyay, N. (2021) How to backup and restore MySQL databases using the mysqldump command. https://www.sqlshack.com/how-to-backup-and-restore-mysql-databases-using-the-mysqldump-command/ (Accessed: December 18, 2023).

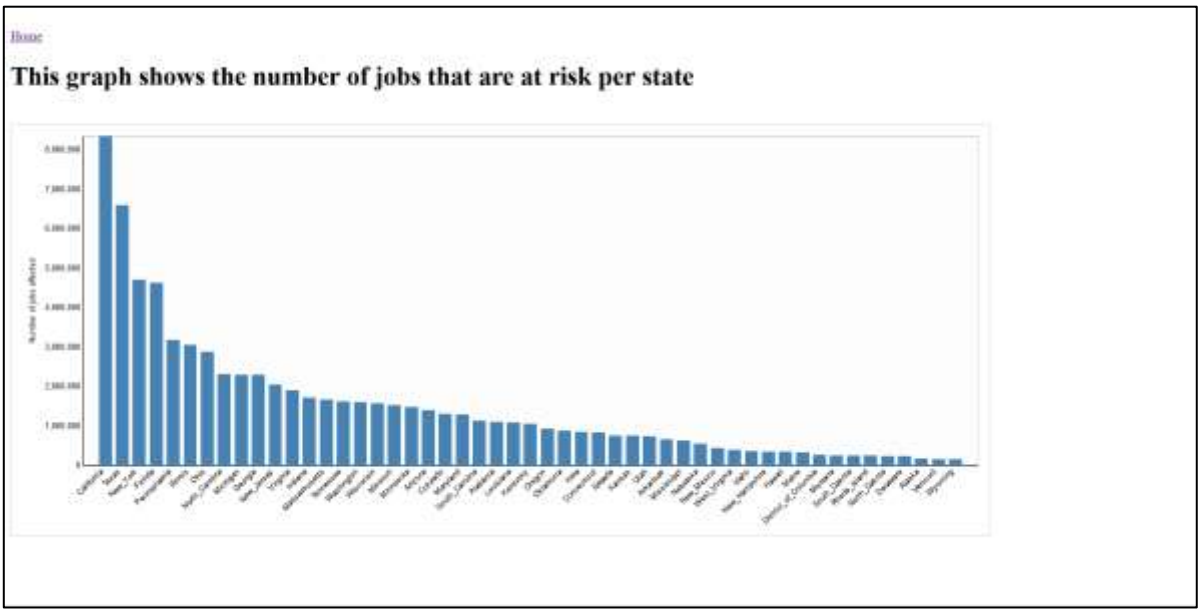
Appendix

Running application screenshots

Home page

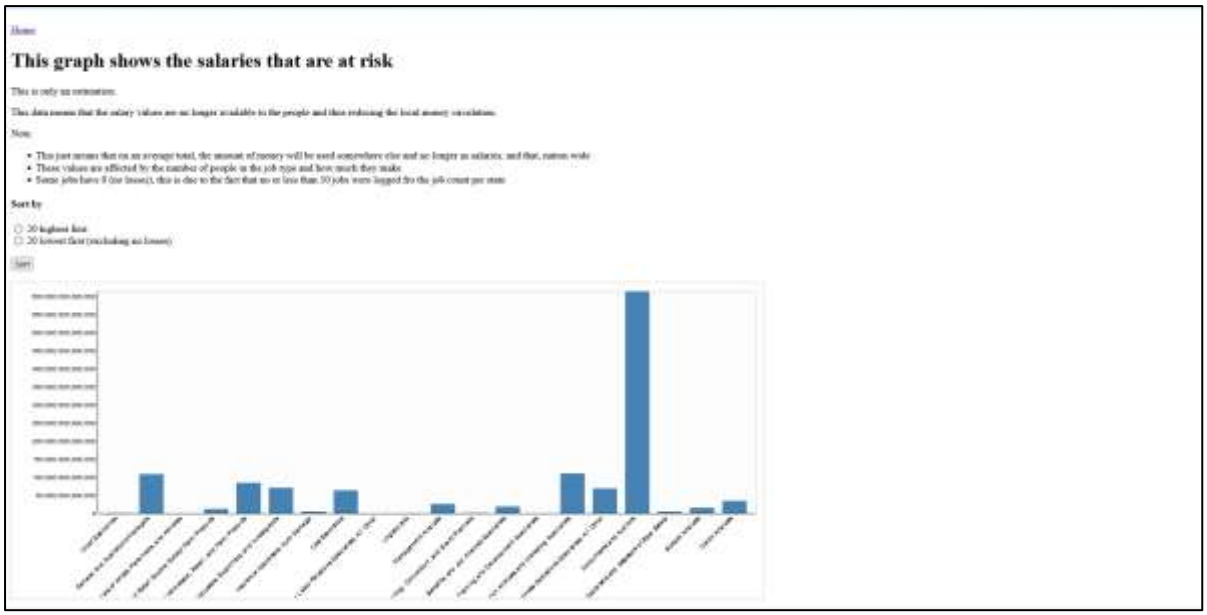


Number of jobs at risk

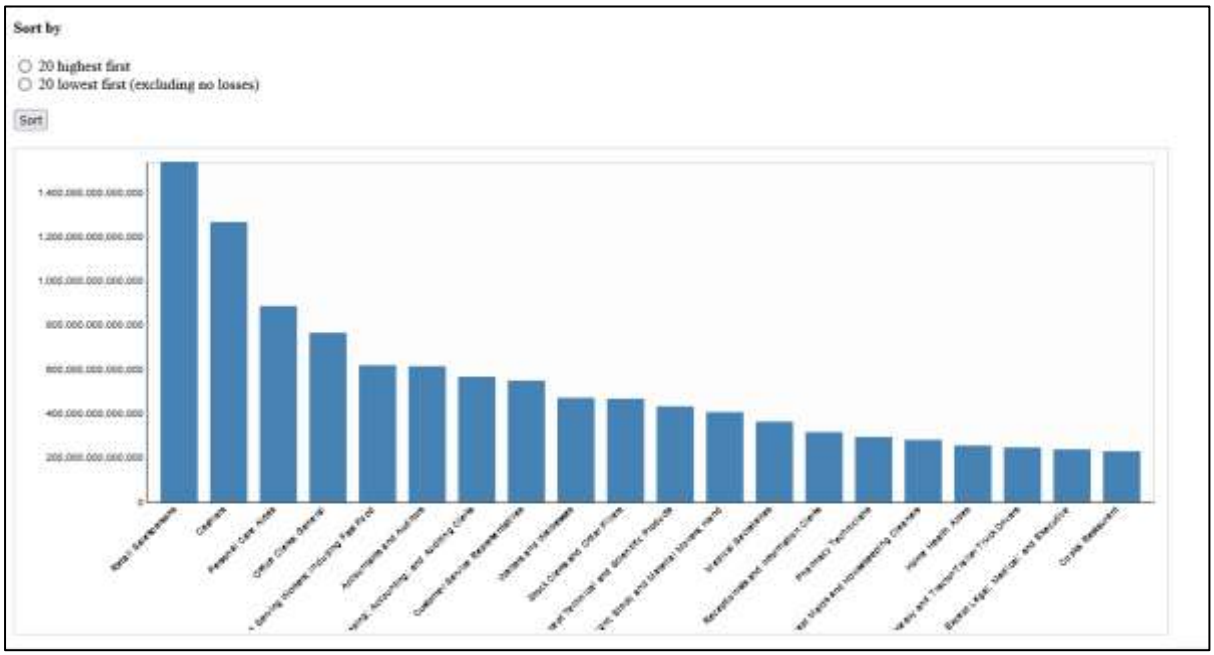


Salaries affected from automation layoffs

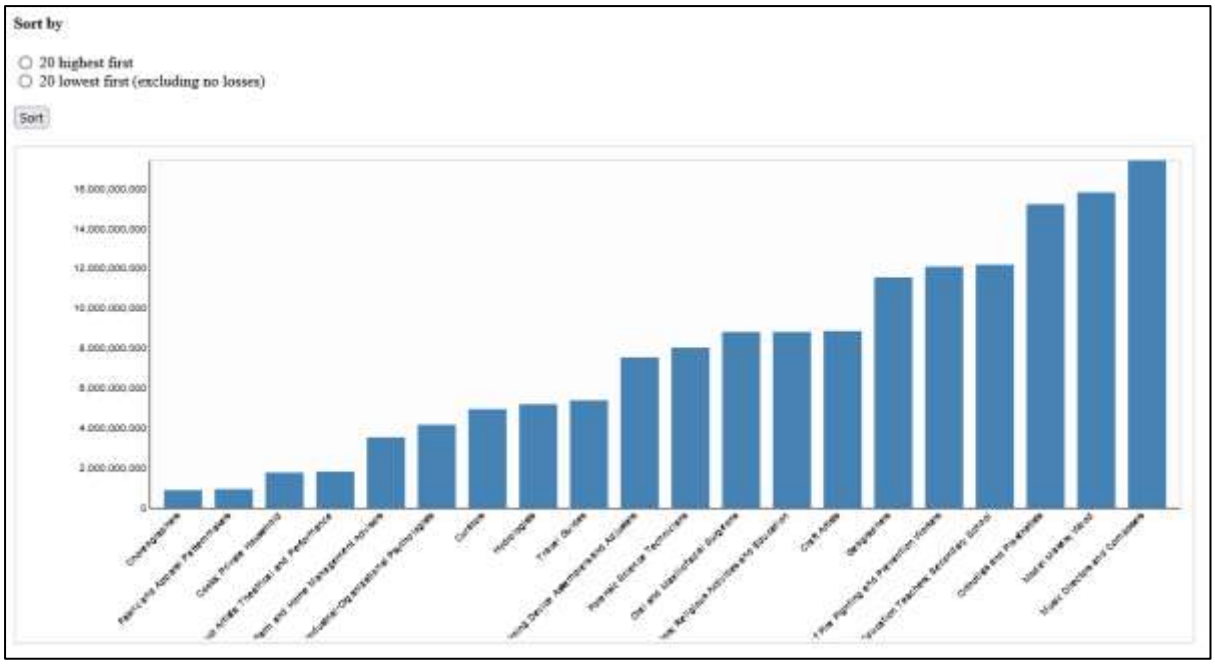
No sorting



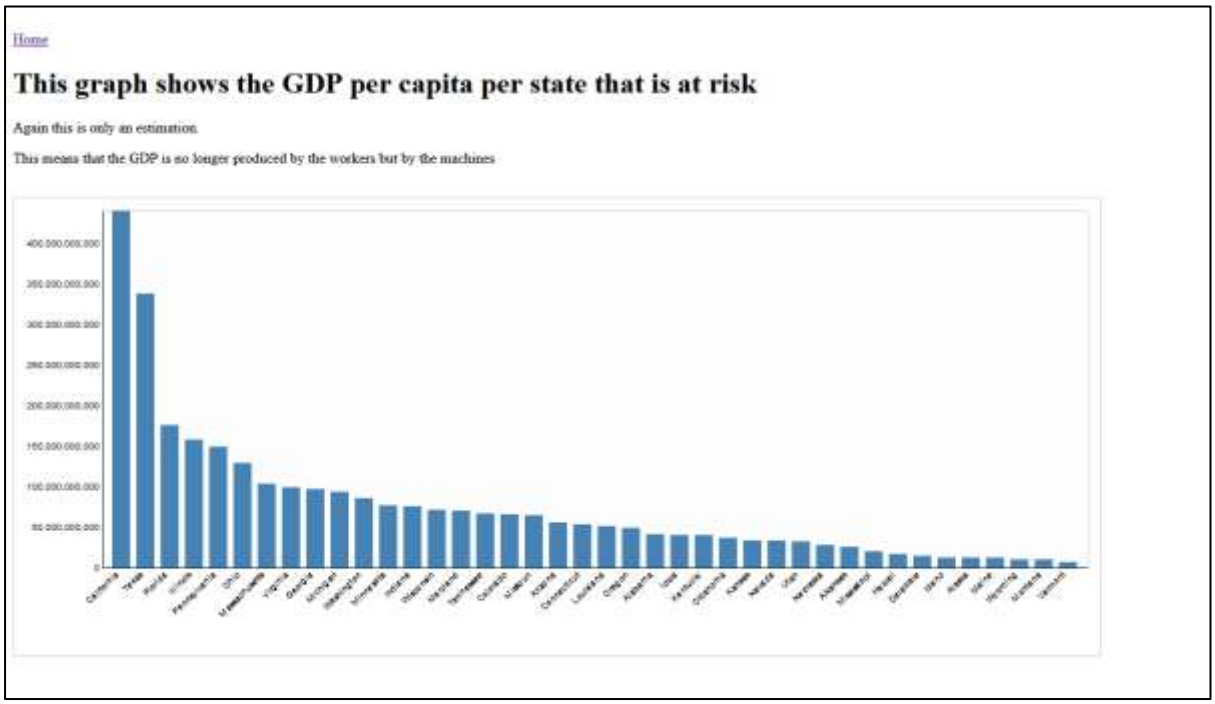
High first



Low first



GDP per capita that will be lost, for all states



GPD per industry that will be lost, filtered for a state