# Report for Comp6234 - Data Visualisation

## Salary distribution and skills required for IT-related jobs in the USA

Junming Zhang
Electronics and Computer Science - Data Science
University of Southampton
Southampton, United Kingdom
jz1g17@soton.ac.uk

*Abstract*

**The aim of Data Visualisation coursework is to tell a data story with produced graphs and supported texts. This document will be divided into 4 main parts. First part will describe the chosen topic and used datasets of this story. Next part will introduce the charts choices and design with concepts covered in the course. Third part will display the produced graphs with flow and evaluation in detail. In the final part, the future improvements and conclusion of this coursework will be illustrated.**

***Keywords-component; IT; salary; skills; USA***

## I. INTRODUCTION (*TOPIC AND DATASETS*)

### *A. Topic of data story*

ECS students may find technology and computer-related jobs after graduation. Therefore, this story will analysis the salary distribution of different jobs and number of job positions in the USA. Moreover, which skills are required to apply IT-related job will be illustrated.

### *B. Datasets choice and processing*

Three datasets are chosen to do the related analysis:

- US jobs on Monser.com
- U.S Technology Jobs on Dice.com
- US state

Two datasets which crawled from job searched websites Monster.com and Dice.com provided lots of information which shown in Table 1 and 2. Job positions, location, salary and skill required are valid data which can be observed after cleaning the original dataset.

Table 1. Header of data from Monster.com

| | | |
|---|---|---|
| country | date added | country code |
| job type | job board | has expired |
| location | organization | page url |
| salary | sector | unique id |

Table 2. Header of data from Dice.com

| | | |
|---|---|---|
| site name | shift | employment type job status |
| company | job id | job location address |
| job title | unique id | job description |
| postdate | skills | advertiser url |

Data visualization contains several steps [1]:

- Raw data collection
- Data processing
- Data clean
- Data analysis
- Data refine
- Data visualizations

For charts choices and design, the original datasets need processing and clean because there are lots of useless and unnormalize data.

Therefore, the first step is choosing key data to analyze the topic of this story which is the salary distribution in the USA. In this case, salary, sector and location in Table 1 are needed to research. However, there are lots of missing and unnormalize data of them which is shown in Fig 1.

| | location | salary | sector |
|---|---|---|---|
| 0 | Madison, WI 53702 | NaN | IT/Software Development |
| 1 | Madison, WI 53708 | NaN | NaN |
| 2 | DePuy Synthes Companies is a member of Johnson... | NaN | NaN |
| 3 | Dixon, CA | NaN | Experienced (Non-Manager) |
| 4 | Camphill, PA | NaN | Project/Program Management |
| 5 | Charlottesville, VA | NaN | Experienced (Non-Manager) |
| 6 | Contact name Tony Zerio | NaN | NaN |
| 7 | Austin, TX 73301 | NaN | Experienced (Non-Manager) |
| 8 | Austin, TX 78746 | NaN | Customer Support/Client Care |
| 9 | Chesterfield, MO | NaN | NaN |
| 10 | Berryville, VA 22611 | NaN | Customer Support/Client Care |
| 11 | Columbus, IN | NaN | Customer Support/Client Care |
| 12 | Boston, MA | NaN | NaN |
| 13 | Houston, TX 77098 | 9.00 - 13.00 $ /hour | Entry Level |
| 14 | Houston, TX | 80,000.00 - 95,000.00 $ /year | Building Construction/Skilled Trades |

Fig. 1. 15 rows data in job from Monster.com

The second step is data clean. Remove the row contains NaN value first, and then normalize the data in salary. It could found 2 main types of salary, hourly and yearly wage. Therefore, for an hourly wage, calculate the mean value of two

number on both side of "- "and multiply by 2080.6 to obtain yearly wage. For yearly wage, remove the "," in numbers and calculate the mean value of it.

The location information contains the name of cities and abbreviated name of states in the USA. Therefore, the last dataset is related to whole name and abbreviated name of states in the USA. Therefore, the final data can be classified into states and cities. Fig 2 and 3 are the cleaned data.

| | location | salary | sector |
|---|---|---|---|
| 0 | Florida | 24967.200 | Customer Support/Client Care |
| 1 | Colorado | 147722.600 | IT/Software Development |
| 2 | California | 35370.200 | Administrative/Clerical |
| 3 | Minnesota | 37450.800 | Accounting/Finance/Insurance |
| 4 | Minnesota | 35370.200 | Business/Strategic Management |
| 5 | Nevada | 39531.400 | Experienced (Non-Manager) |
| 6 | California | 83224.000 | Experienced (Non-Manager) |
| 7 | California | 20806.000 | Accounting/Finance/Insurance |
| 8 | Massachusetts | 40571.700 | Accounting/Finance/Insurance |
| 9 | Michigan | 34329.900 | Accounting/Finance/Insurance |

Fig. 2. 10 rows data in cleaned state-salary-sector

| | location | salary | sector |
|---|---|---|---|
| 0 | Houston | 22886.492622 | Entry Level |
| 1 | Denver | 31729.001135 | Experienced (Non-Manager) |
| 2 | Clutch City | 119633.938706 | Experienced (Non-Manager) |
| 3 | Willoughby Hills | 40571.509648 | Experienced (Non-Manager) |
| 4 | Telecommute | 98828.036322 | Experienced (Non-Manager) |
| 5 | Margate | 24967.082860 | Customer Support/Client Care |
| 6 | Denver | 147721.906924 | IT/Software Development |
| 7 | Anaheim | 35370.034052 | Administrative/Clerical |
| 8 | Eden Prairie | 37450.624291 | Accounting/Finance/Insurance |
| 9 | Minneapolis | 35370.034052 | Business/Strategic Management |

Fig. 3. 10 rows data in cleaned city-salary-sector

To find the required skills in IT related jobs. Match IT and computer related data by regular expression and split those skill with "," to obtain the individual skill. The last result is shown in Fig 4.

| | |
|---|---|
| 219548 | XML |
| 219549 | HTML |
| 219550 | Python |
| 219551 | Ruby |
| 219552 | Go |
| 219553 | Clojure |
| 219554 | Java |
| 219555 | NoSQL-Database |
| 219556 | SQL |
| 219557 | JaveScript |

Fig. 4. 10 rows data in cleaned skill required

## II. CHARTS CHOICES AND DESIGN

Right charts choose can lead a clear result to the audience. Therefore, the choice and design of charts are significant.

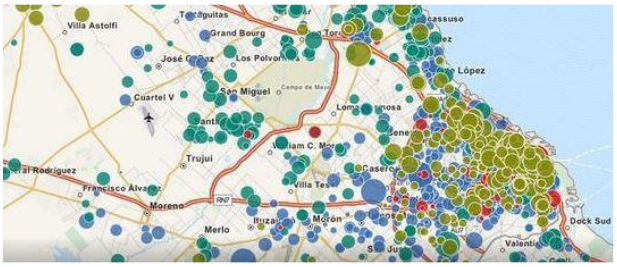


Fig. 5. Map chart

The aim of this story is telling people the distribution of salary in the USA. Therefore, the average salary of different states should be displayed. In this case, a map chart which shown in Fig 5 could be used to display and analysis salary of different states.

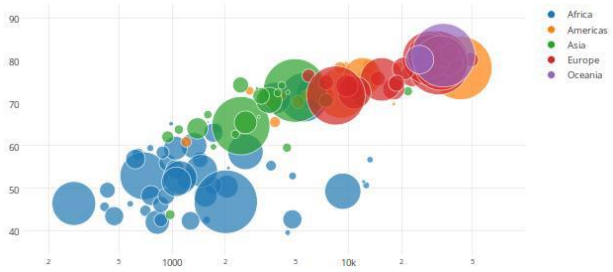


Fig. 6. Bubble chart

Farther more, the salary comparison of different jobs can be displayed in bubble chart which is shown in Fig 6. In this chart, the numerical value can be compared to the height of vertical coordinates. At the same time, the number of different jobs offered in each city can be compared to the size of bubbles.


Fig. 7. Cloud words chart

The purpose of cloud words which shown in Fig 7 is to statistical and analysis the most useful skill in IT related jobs. It could count the frequency of different skills in Fig4. Therefore, students could have a direction to learn the beneficial skill.

## III. GRAPHS IN DETAIL

3 graphical are produced to describe the salary and skills required for jobs in the USA. First two diagram is generated by Tableau and the last diagram is generated by cloud words creator.

A good graphical display based on several features [2]:

1. No need to understand a subject by hand.
2. No data distorting.
3. Combine large datasets.
4. Compare data by eyes.
5. Display data on different levels.
6. Integrate datasets with statistical and text description.

Therefore, all produced diagrams are referenced to these rules.

### *A. First Diagram*

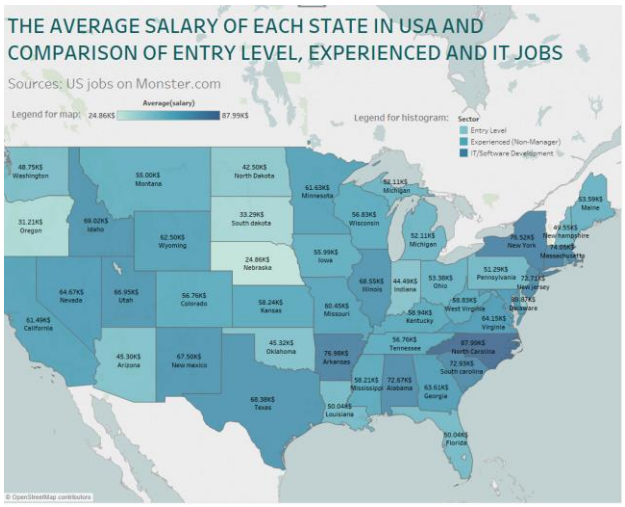


Fig. 8. First part of produced diagram 1: map chart

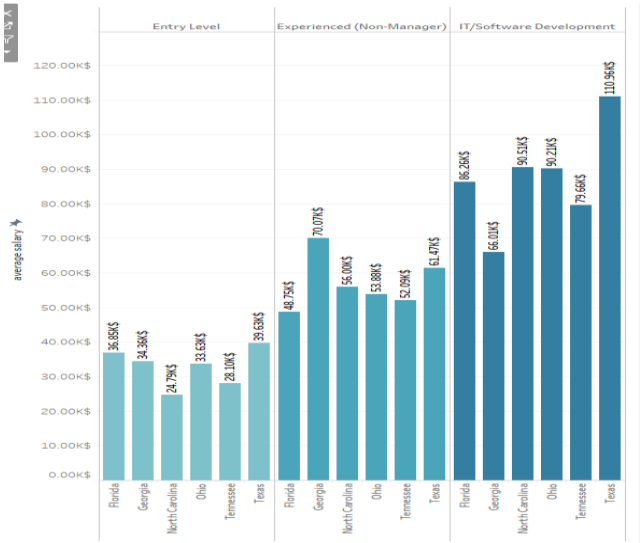


Fig. 9. Second part of produced diagram 1: histogram chart

The aim of this part is to provide a general salary distribution situation in the USA and compare IT relevant jobs with other types of jobs. Therefore, the first diagram is consisting of map chart and histogram chart which are shown in Fig 8 and 9.

Map chart is produced by Tableau by using the processed data "state_salary_sector.csv". Statistics and calculate the average salary in each state and plot the result to the map of USA. The degree of colour is used to represent the number of salaries. According to the 12-colour palette adapted for colour blindness, Shades of cyan is chosen for the main clour for this chart[3].

The distribution of average salary in each state is labelled in Fig 8 now, people could easily observe the result. According to Bureau of Labor Statistics report for the United States, there is 41.52K$ yearly median income of a person for the full-time worker in 2017. The result from map chart indicates the salary of most states are varied at 40~55K$ which is similar to the official statistics which means this chart is evidence readable. At the same time, the high salary state can be observed by deep colour area which is North Carolina, New York and Arkansas, the low salary state is Nebraska.

The advantage of map chart is the good visual, people could find the high salary states visualized, at the same time, people could choose the workplace by analysis the geographic position.

Entry level and experienced jobs are used to compare the salary of IT-relevant jobs. Using Tableau to statistics these three types jobs, 6 states contains salary data of these three types job at the same time. Therefore, the histogram chart could be generated. Using the same clour to represent one type job and divided them by straight line could let people have a better visual.

The advantage of the histogram is it provides a clear visual to do compassion. It could find the salary of IT/Software development jobs are higher than experienced and entry-level jobs which means the IT-relevant job is a good choice at present.

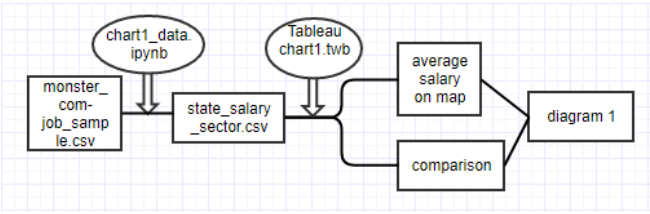


Fig. 10. Flowchart of first diagram: ipynb and twb files will be submitted

This two chart will be combined together with the flow shown in Fig 10 and display in the HTML page with the reason why different salary distribution in different states and type of jobs will be analysed.

## B. *Second Diagram*

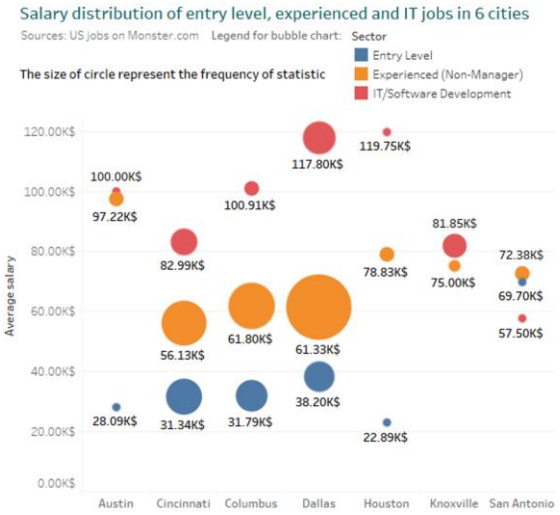


Fig. 11. Produced diagram 2: bubble chart

In the first diagram, people already checked and compared the average salary of different type of jobs. Therefore this part aims to provide a new type of chart to compare the number of offered positions. Moreover, it focuses on different cities now. There are 3 variables, therefore, the bubble chart which shown in Fig 11 is produced.

The advantage of the bubble chart is it could be used to display three dimensions data. Therefore, the comparison of three type jobs and the relationship between job positions can be displayed in the same diagram.

This diagram is generated by Tableau by using the processed data "city_salary_sector.csv". According to statistics, 7 cities contains salary data of entry-level, experienced and IT/Software development jobs. Then, the average salary of them can be calculated and plotted to the bubble chart which the size of the bubble can be used to represent the offered number of positions. The general flow also is shown in Fig 12.

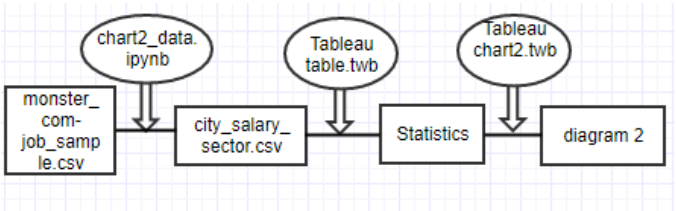


Fig. 12. Flowchart of second diagram: ipynb and twb files will be submitted

According to Fig 11, it could find in collected data, only 7 cities from Texas and Ohio offered 3 types jobs at same time. It similar to result in part 1, IT-relevant jobs have a higher salary but the job positions of IT jobs are less than other jobs because the knowledge of IT jobs is higher than other jobs. Therefore, which skills provide more opportunities to find IT jobs will be analyzed in part 3.

## C. *Third Diagram*


Fig. 13. Produced diagram 3: cloud words chart

This part aims to provide a new type of chart to investigate the most useful skills in IT-relevant jobs. The frequency of different skills in employment website can be used to show the popular skills of IT jobs at present. Therefore, the cloud words chart which shown in Fig 13 is produced.

The advantage of cloud words chart is high-frequency words are display clearly, in this case, people could find which skills are useful for finding IT-relevant jobs.

This diagram is generated by cloud words tool by using the processed data "skill_output.csv". The diagram and frequency record file will be generated. The top 10 popular skills are shown in Fig 14.

| JAVA | SOL | JS | LINUX | ORACLE | AGILE | WEB | HTML | CSS | PYTHON |
|---|---|---|---|---|---|---|---|---|---|
| 3148 | 2809 | 1746 | 1409 | 1288 | 1202 | 1158 | 1072 | 929 | 928 |

Fig. 14. Rank of most used skills in IT jobs

The general flow is shown in Fig 15.

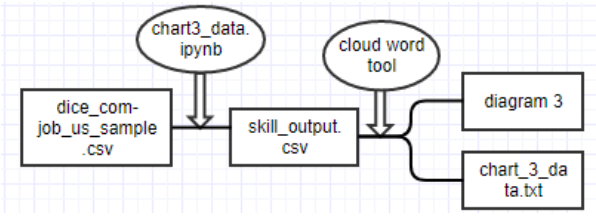


Fig. 15. Flowchart of third diagram: ipynb and text files will be submitted

It could find people who know Java, JavaScript, database and front-end development relevant have more opportunities to find a job. In addition, Agile development and Python become more and more popular in recent years.

## IV. FUTURE IMPROVEMENTS AND CONCLUSION

### A. *Improvements*

The initial plan of this story is introducing the salary distribution of states in the USA and find the best city for IT-relevant job. However, the dataset from employment website is a sample version, the completed data is rechargeable, therefore, 22000 rows of data cannot cover all cities in the USA. Moreover, only hundreds of available data remain after data clean. Table 3 record the usable data in processed data.

Table 3. the number of processed data which contains salary data

| Experienced | manager | accounting | Entry level | sales | IT |
|---|---|---|---|---|---|
| 583 | 268 | 222 | 153 | 108 | 76 |

In this case, it is hard to find which city is most suitable for IT jobs. Therefore, this story transferred to compare entry level, experienced and IT jobs.

Thus, the key of future improvement is collect more data. The rechargeable data cost 20$ which could provide 4.7 million job lists. According to present ratio, there will be 0.4 million useable data. Another way is crawler data from the employment website. With plenty of data, the statistic result will be more accurate.

For the map chart, it can be improved by displayed more statistic elements which like median, variance and quantity, these variables can be shown in the hidden layer, once people click one state in the map, relevant information will be displayed. D3 could be used in this part to plot a visible diagram.

For the histogram chart, existing 6 states are those states which contain salary data for entry-level, experienced and IT jobs. Therefore, it could be extended or replaced by well-developed states in the USA. The audience could obtain more information in this chart.

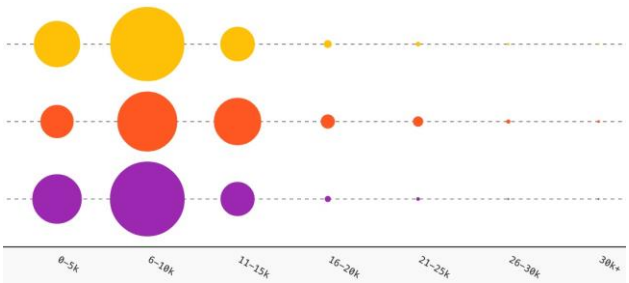


Fig. 16. Initial plan of bubble chart

For the bubble chart, it has more potential to show more information because the initial plan of this chart is to display the salary of different job in different ranges which are shown in Fig 16. The size of the bubble can be used to describe the number of salary sections. Thus, people could know the salary distribution of different job. However, the data of IT relevant jobs is not enough to draw this chart. In this case, this part changed to display the number of positions. A salary distribution chart can be plotted in the further improvement.

### B. *Conclusion*

The aim of this coursework is telling an interesting data story to the audience by using graphics and supported text.

The structural of this story and chart design are good because it introduces the salary distribution in the USA by different levels. First the average salary in each state and comparison of entry-level, experienced and IT jobs are described in the first diagram. At the same time, the analysis of the reason why some states have a higher salary and some states have lower salary by the geographic position and development history. After that, the position of different type of jobs and salary comparison in different cities are introduced by the second diagram. Finally, the useful skills to find an IT relevant jobs are listed by the third diagram. The structural is in descending order, the audience could obtain information layer by layer.

The result of this story is Nebraska, South Dakota and Oregon have a lower average salary. North Carolina, Arkansas and New York have a higher average salary. The salary for IT relevant jobs is much higher than jobs offered to entry level and experienced level, but the number of IT positions is much less than other jobs. A job seeker who know Java, JavaScript, database and relevant front-end skills have more opportunities to find a job. In addition, Agile development and Python become more and more popular in recent years.

## REFERENCES

[1] S. Elena, Introduction to data visualisation, University of Southampton, 2017 October 03 [online]. Available: http://edshare.soton.ac.uk/18967/ (accessed 27th December 2017)

[2] T. Edward, The Visual Display of Quantitative Information, 1983.

[3] S. Elena, Visual perception and information design for the mind - I, University of Southampton, 2017 November 14 [online]. Available: http://edshare.soton.ac.uk/19126/ (accessed 27th December 2017)