Salary of different degrees

Khadija Motlekar | Data Visualisation and the Web | 14/12/18

Introduction:

In this report, I will analyse different degrees and the salary received by the degrees. I will take into account the starting salary, mid-career salary and the field of the degree.

Research topic:

The motivation for this topic came from an article I read about on salary and degrees [1]. After reading the article, I wanted to know which degree pays the highest salary at the beginning and midway through a person's career. As some degrees have a higher salary compared to others. These salaries are divided into starting salary and mid-career salary. The change in the salary from start to mid-career is high for some degrees and low for others. I wanted to find out whether different fields such as; science, arts, commerce, engineering have a higher salary that others. In addition to this, I was curious about finding out which degree has the highest and lowest salary difference between the starting salary and mid-career salary.

In this report, I want to find out:

1. if there is a significant change in the starting and mid-career salary of the degrees
2. does the type of field have an effect on the starting salary
3. which degree has the highest and lowest difference in salary

From the many articles I read, the information that was highlighted the most was that science degrees, earn more than any other field. The highest paying degree is a science degree; medicine, veterinary and dentistry, in second place is Economics an Arts degree [2]. When it comes to choosing a career, people mostly pick the most generic ones like law, psychology or medicine (in 2017), without actually knowing how much these degrees earn. From this research, I am hoping to point out the degrees that have the highest and lowest starting salary as well as mid-career salary.

Data:

The data on degrees and the salary they earn was taken from a website called Kaggle [3]. Kaggle is an online community for data scientists and machine learners. Kaggle allows users to find and publish datasets, explore and build models for predicting and describing the data. I came across this dataset while looking for a dataset for a previous idea of mines.

The dataset contained 8 columns and 50 rows. The dataset contained columns for degrees, starting median salary, mid-career median salary, percent change from starting to mid-career salary, mid-career 10^{th} percentile salary, mid-career 25^{th} percentile salary,

mid-career 75th percentile salary and mid-career 90th salary in dollars($). I was not required to change the file type of the data as the file was a comma-separated values (CSV) file. All of the columns in the dataset were useful and necessary for the research I wanted to conduct. Due to this, I didn't drop any of the columns in the code. Along with the columns given, I added my own column to describe the field in which the degrees belonged to [4]. The fields columns was an addition made by me to further research the difference in salary based on the field in which the degrees belong in, for example Biology is a Science degree and Finance is a Commerce degree.

The data given was clean and tidy, there was no need to drop any NaN values from the data. I renamed the column names to short specific names as the given names where pretty long to type and read. In addition to this, I was required to change the data types of the columns from objects to more specific types for visualisation. Almost all of the columns were objects except the change in percent column which was *float64*. For the data to be suitable for visualisation, I had to change the degree column and the field column into *category* from *object* and change startSalary, midSalary, mid10, mid25, mid75 and mid90 to *float64* from *object*. However before I could change the data types, I had to remove the $ sign and commas from the data as pandas kept classifying the columns with the $ sign and commas as a *string*. To remove the $ sign and commas, I created a 'for loop' to go through the data and replace the $ sign and commas with nothing.

startSalary	startSalary
$46,000.00	46000.0
$57,700.00	57700.0
$42,600.00	42600.0
$36,800.00	36800.0
$41,600.00	41600.0

Figure 1 the dollar sign and commas removed from the column

```
Undergraduate Major                                  object
Field                                                object
Starting Median Salary                               object
Mid-Career Median Salary                             object
Percent change from Starting to Mid-Career Salary    float64
Mid-Career 10th Percentile Salary                    object
Mid-Career 25th Percentile Salary                    object
Mid-Career 75th Percentile Salary                    object
Mid-Career 90th Percentile Salary                    object
dtype: object
```

```
degrees       category
field         category
startSalary    float64
midSalary      float64
change         float64
mid10          float64
mid25          float64
mid75          float64
mid90          float64
dtype: object
```

Figure 2 Changes made to the data types of the columns and the names of the columns

Exploratory and explanatory data visualization:

Most of the variables in this dataset are **numerical- ratio**, with a couple that are **categorical- nominal**. There aren't any variables with the data type interval and ordinal as the difference between the salaries varies for each degree and none of the categorical data is ranked in order. This dataset looks at 50 undergraduate degrees from different fields and the salary each degree earns at the start of the career and midway through the career. The salary given is the median amount of salary a degree earns. The mid-career salary is further divided into four sub categories; mid-career 10th percentile, mid-career 25th percentile, mid-career 75th and mid-career 90th salary. This dataset looks at how the salary of a particular degree changes over time.

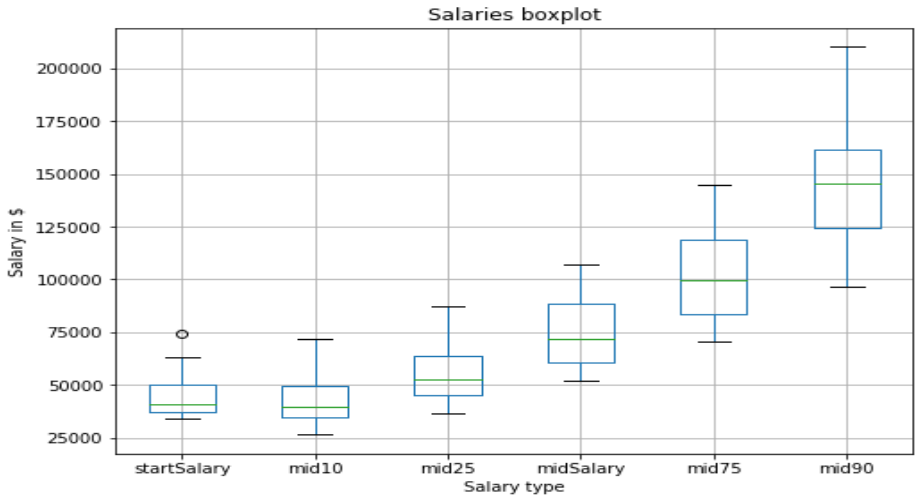
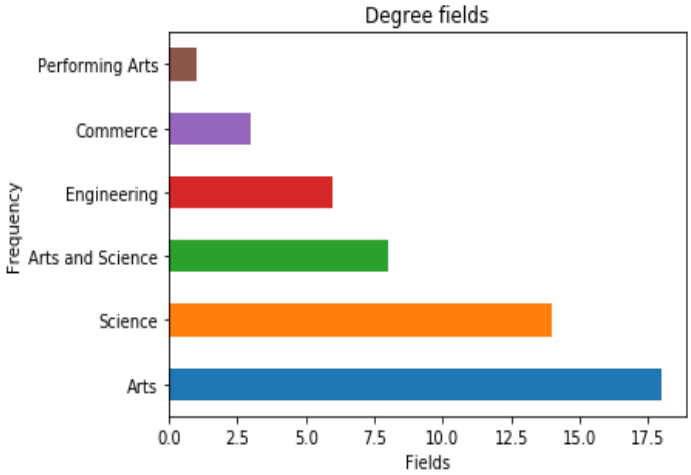


Figure 3 boxplot representing the salaries over time

From the boxplot shown, I can see that there is a significant change in the salary over time. There is an outlier present in the startSalary of the degrees. When I compare start salary and mid10 salary, I can see that the lowest salary for mid10 is smaller than the lowest salary for startSalary. This could suggest that the salary for a particular degree decreases for a small period of time. The change between the salaries is very gradual and almost linear. The lowest salary goes from $25000 to around $95000 and the highest goes from $75000 to around $210000. I chose a boxplot to visualise this data because boxplots make it easier and quicker to identify the minimum value, maximum value, mean, 1st quartile and the 3rd quartile. Using this information, the reader can then calculate the Inter Quartile Range (IQR) for each of the box plots. The data type for the salaries is **numerical- ratio** type as salaries there a clear definition of 0.

From this bar chart I can see that most of the degrees from this dataset are from the Arts field with the frequency of 18 and only a small number of the degrees are from the




```
Arts                 18
Science              14
Arts and Science      8
Engineering           6
Commerce              3
Performing Arts       1
Name: field, dtype: int64
```

Figure 4 horizontal barchart to represent the different fields of the degrees

Preforming Arts field with the frequency of 1. The data type for fields is **categorical-nominal** type as the data is discrete, in no natural order and in name only. I chose a bar chart to visualise the data because bar charts displays the highest and lowest data in order of frequency. This allows the reader to identify important information quickly and easily.

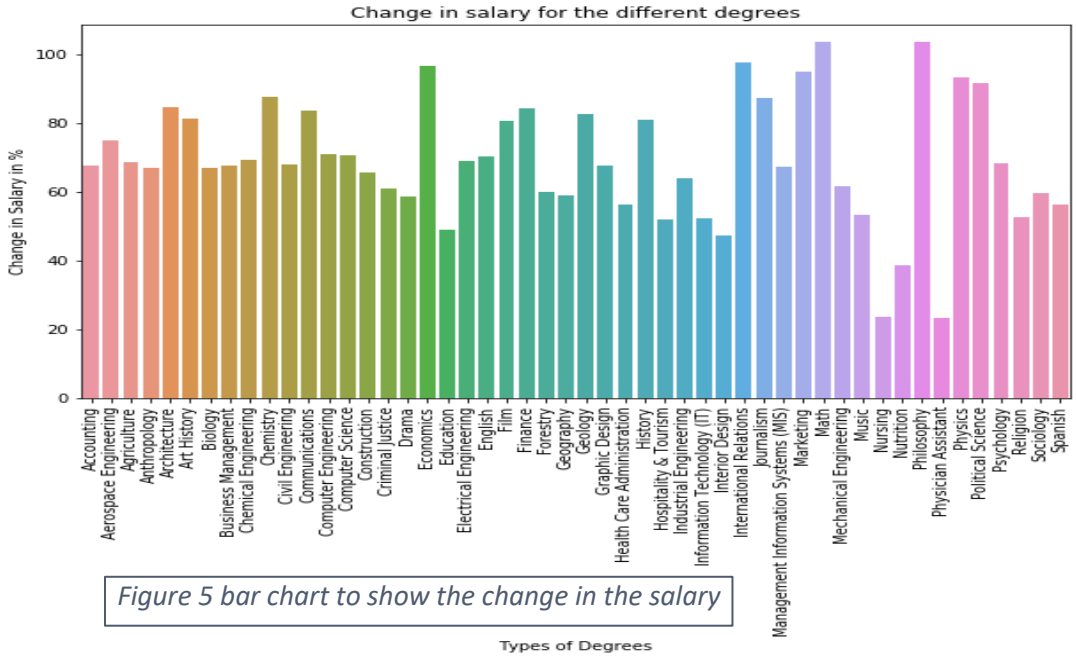


Figure 5 bar chart to show the change in the salary

This bar chart shows the change between startSalary and midSalary in percent. The highest change in salary is for the degrees Maths and Philosophy and the lowest change in salary is for Nursing and Physician Assistant. The data type for the change is **numerical-ratio** type as there is a clear definition of 0 in the data set. The data type for the different degrees is **categorical- nominal** type as the degrees don't relate to each other in any particular way.

Figure 6 grouped bar chart of start salary and mid-career salary

This grouped bar chart shows the startSalary and the midSalary of the degrees. The startSalary is showed by the blue bar and the midSalary is shown by the orange bar in the grouped bar chart. The data type for startSalary and midSalary is **numerical- ratio** type as there is a meaning for 0. From the chart, I can see that for most degrees, the midSalary is increased by 25% compared to startSalary. Moreover, there are some degrees, with a raise that is minor for example nursing. I chose a grouped bar chart to display this data because a grouped bar chart allows us to see both the starting and mid-career salary side by side to each other and to understand the difference between them.

Conclusion:

In conclusion, the change between the startSalary and midSalary of certain degrees is greater than other salary. There are some salaries where the midSalary is twice of the startSalary while other have a 25% increase to it. In addition to this the field in which the degrees belong in had little effect on the starting salary. However, from looking at the graphs, it can be said that mostly engineering degrees have a higher startSalary. From the charts, it was pointed out that Maths and Philosophy had the highest difference and Nursing and Physician Assistant had the lowest difference.

During pre-processing the data, I learned that it is crucial that the data is in the correct data type for accurate visualisation to take place. I ran into a problem due to the data types of my data not be correct and so I had to change the current data types of the variables by removing any information that was overriding the code. In this data, I had to remove the $ dollar and commas from the data in order to assign the data type as *float64* from *object*. For next time, I will find data that is in pounds (£) rather than dollars as it would have been helpful to see which degree earns the most and least in the UK.

Reference:

[1] https://www.independent.co.uk/news/business/news/the-10-best-degree-subjects-if-you-want-to-make-a-lot-of-money-after-university-a6981951.html - Article (accessed on 07/12/18)

[2] https://www.bbc.co.uk/news/education-41693230 - (accessed on 07/12/18)

[3] https://www.kaggle.com/cdelany7/exploration-of-college-salaries-by-major - (accessed on 07/12/18)

[4] https://www.otago.ac.nz/courses/otago036839.html - (accessed on 08/12/18)

[5] http://queirozf.com/entries/pandas-dataframe-plot-examples-with-matplotlib-pyplot#plot-column-values-as-a-bar-plot – used as reference to plot some of the graphs (accessed on 09/12/18)