Data science

Tools & techniques

**Name:** Teun Vlassak

**Datum:**  15-01-2023

**Titel:**  Data science

**Ondertitel:** Tools & techniques

**Naam studenten:**  Teun Vlassak

**Opleiding:**  Logistiek Management FEM, voltijd

                                                      Hogeschool Arnhem Nijmegen

                                                         Laan van Scheut 10

                                                         6526 EM Nijmegen

                                                        T: 024 – 353 05 00

**Klas:**  MDD Minor data science

**Docentbegeleiders:**  Chris Scholten

**Datum:** 06-01-2023

**Module:** MINOR

**Kans:** 1

Verklaring eigen werk

The students below declare:

* that everything described in this paper is own work, and that by means of source, it is indicated which parts have been copied
* that nothing from this paper has been or will be passed on to others in the future.

|  |  |
| --- | --- |
| Class and group | **MDD Data science** |
| Teacher | **Chris Scholten** |
| Name and paper | **Data science Tools & techniques** |
| Dat of paper | **06-01-2023** |

|  |  |  |
| --- | --- | --- |
| Name | Studentnumber | Sign |
| **Teun Vlassak** | **623435** |  |

Executive summary

An executive summary would be that understanding the salaries of data scientists is important for both individuals considering a career in this field and for employers looking to set competitive salaries.

I used the process of data preparation, including techniques such as counting missing values, handling duplicate observations, and converting currency, is crucial for preparing the data for analysis.

The CRISP-DM model is used to make predictions or identify patterns in the data. By summarizing the salary data to gain insights into the range and distribution of salaries for data scientists. I identified patterns to compare with other professions and identify factors associated with higher or lower salaries. However, it's important to keep in mind that these predictions will be based on statistical models and may not be completely accurate, and that it's important to validate the predictions with real-world data.

The main question is can I predict my future income as data scientist based on the dataset given?   
Although it is important to keep in mind that these predictions will be based on statistical models and may not be completely accurate.

The answer is no, the data I used was from America and the United Kingdom and I can not compare them with a European student. The difference between location, age, and entry salary level is too high to compare.

In conclusion, understanding the salaries of data scientists is important for both individuals considering a career in this field and for employers looking to set competitive salaries. Through various data preprocessing techniques such as counting missing values, handling duplicate observations, and converting currency, the data can be prepared for analysis.

Index

[1.1. Data understanding 5](#_Toc124683084)

[2.1. Data preparation 6](#_Toc124683085)

[3.1. Modelling 8](#_Toc124683086)

[4.1. Conclusion 10](#_Toc124683087)

[5.1. Evaluation 10](#_Toc124683088)

# 1.1. Data understanding

The focus is on understanding the business objectives and context for the project, as well as identifying the questions that the project is intended to answer.

With this project I want to have an understanding of my upcoming salary after this study. It is difficult to predict my future salary as a data scientist because it is depending on different variables;  
1. Job\_title: Data scientist, data engineer etc.  
2. Job\_type: Full time, part time, internship  
3. Experience level: Senior, mid & entry  
4. Location: New York, boston, London.  
5. Salary currency: USD, Euro & GBP.  
6. Salary: -  
7 Remote location Yes or no.

Also it is important to understand the demand of the job market at the time you are seeking employment.

The reason why my dataset is good enough to do the research for salary level of data scientist;   
1. Completeness: My dataset is complete with many different variables like Job title,   
 job type and experience level.  
2. Accuracy The dataset was not without errors but I fixed them with the data cleaning.  
3. Validity The dataset is valid to make a good and complete answer to the question.  
4. Timeliness The dataset is current and relevant to the task.

Main question  
- Can I predict my future income as data scientist based on the dataset given?   
Although it is important to keep in mind that these predictions will be based on statistical models and may not be completely accurate.

# 2.1. Data preparation

Data preparation refers to the process of cleaning, transforming, and organizing raw data so that it can be used for analysis or modelling. I have taken the following steps to create a useable dataset;

**1. The code started with importing the raw datafile as a csv.**Speaks for itself.

**2. “Salary;” had to be numeric and changed to “salary”.**Changing "Salary;" to "salary" ensures consistency and readability of data, and also helps to ensure that the data is recognized as numerical instead of character.

**3. Count the missing values to know if the data is complete.**Counting missing values in a data set can help determine if the data is complete and ready for analysis or modeling. It also allows identifying patterns in missing data, so that appropriate strategy can be chosen to handle missing data.

**4. Count total duplicate observation to know if the dataset is complete.**Counting duplicate observations can help determine if the data is complete and accurate, and also helps to identify and remove unnecessary duplicates, ensuring that the final dataset is a representative sample of the population.

**5. Look at the names if they are complete and correct.**Speaks for itself.

**6. I need to get a count of the number of observations of each unique value in the “job\_title”. The** **missing values will not be included.**Counting the number of observations of each unique value in the "job\_title" column can provide insights into the distribution of job titles in the data set, which can be useful for identifying patterns, trends, and for making informed decisions. Excluding missing values in the count will give a more accurate representation of the data.

**7. Look at the values of currency.**To look if they are different.

**8. That some of the values in the salary\_currency column are not in US dollars (USD), and that all of the values need to be converted to USD.**Converting the values in the salary\_currency column to USD is important for accurate and consistent comparison of salary values across the data set, for making the data comparable to other data sets or industry standards, and for any analysis or modeling being performed on the data.

**9. change intial salary column to salary\_usd\_gbp\_eur and salary\_usd to salary.**Changing the initial salary column to "salary\_usd\_gbp\_eur" and "salary\_usd" to "salary" can be done to make the column names more clear, readable and indicating the currency that is used in the salary column. It also makes it easier to reference the columns in code or analysis, and avoid errors.

**After these data preparation codes I am happy with the end result:**

|  |
| --- |
| Figure 1. |
|  |
| The picture is directly from the R script. |

**After these lines of code I had a useable data package to answer the main question;**- Can I predict my future income as data scientist based on the dataset given?   
Although it is important to keep in mind that these predictions will be based on statistical models and may not be completely accurate.

It is possible to use data science techniques and a given dataset to make predictions about future income as a data scientist, however, it is important to understand that these predictions will be based on statistical models and may not be completely accurate.

Factors such as the quality of the dataset, the chosen model, and the specific context of the prediction (e.g. the job market and economy at the time) can all affect the accuracy of the prediction. Additionally.

3.1. Modelling  
The goal of this phase is to develop a statistical or machine learning model that can be used to make predictions or identify patterns in the data. This phase includes selecting a model, training it on a portion of the data, and testing it on a different portion of the data. The best performing model will be selected based on an evaluation metric. The model can be refined or different model can be selected if the initial model does not perform well. I have made these steps;

**1. summarize the salaries of Data scientists**Summarizing the salaries of data scientists can provide insights on the range and distribution of salaries for the profession, by calculating summary statistics. This can be useful for career considerations and setting competitive salaries for employers.

**2. show on a density plot**The density plot can be a useful visual tool for understanding the data and for making informed decisions based on the data.

|  |
| --- |
|  |
|  |
|  |

**3. Average salary by job title**The resulting salary\_by\_job object is a data frame that contains the "job\_title", "count", "min\_salary", "max\_salary", "mean\_salary", and "median\_salary" for each job title in the ds data frame.

**4. Average salary by experience.**The resulting salary\_by\_experienc**e** object is a data frame that contains the "job\_title", "count", "min\_salary", "max\_salary", "mean\_salary", and "median\_salary" for each job title in the ds data frame.

**5. Average salary by job type.**The resulting salary\_by**\_ job type** object is a data frame that contains the "job\_title", "count", "min\_salary", "max\_salary", "mean\_salary", and "median\_salary" for each job title in the ds data frame.

**6. Qqnorm and Shapiro.test**The purpose of performing the qqnorm and shapiro.test is to assess whether the salary column in the ds data frame follows a normal distribution.

|  |
| --- |
|  |
|  |
|  |

**7. Generalized lineair model (glm)**It looks like you are using the "glm" function to fit a generalized linear model (GLM) to the ds data frame, and then using the summary function to display a summary of the model.

# 4.1. Conclusion

In conclusion, understanding the salaries of data scientists is important for both individuals considering a career in this field and for employers looking to set competitive salaries. Through various data preprocessing techniques such as counting missing values, handling duplicate observations, and converting currency, the data can be prepared for analysis.

Using the CRISP-DM model, statistical or machine learning models can be developed to make predictions or identify patterns in the data. By summarizing the salary data and calculating summary statistics such as mean, median, and standard deviation, it is possible to gain valuable insights into the range and distribution of salaries for data scientists, identify patterns or trends, compare with other professions and identify factors associated with higher or lower salaries.

However, it is important to keep in mind that these predictions will be based on statistical models and may not be completely accurate, and that it's important to validate the predictions with real-world data.

Eventually the main question was if i can predict my salary according to this data.frame.

The answer is no, i can not. The data I used was according to american statistics and you can not compare them with a european student. The difference between location, age, and entry salary level is way to high to compare.

# 5.1. Evaluation

I am very satisfied with the results but after evaluating my script with other script I could have done one thing different and this was the dummy variables.

Dummy variables can be useful in some cases as they allow to include categorical variables in regression models and other statistical models that typically require only numerical input. In some cases, creating dummy variables can make the data preparation process more efficient than other techniques such as converting categorical variables to numerical variables.

With my code the dummy variable did not work well with the collinearity. This made understanding the data and process of modelling more difficult.

# bibliography

1. Cody Baldwin. (2021a, September 9). *Introduction to the CRISP-DM Methodology (Analytics & Data Science)* [Video]. YouTube. https://www.youtube.com/watch?v=q\_okDS2RtzY

*2. Fitting Generalized Linear Mixed-Effects Models*. (2022a, November 1). https://rdrr.io/cran/lme4/man/glmer.html

3. Grolemund, G., & Wickham, H. (2017b). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (1st ed.). O’Reilly Media.

4. Hotz, N. (2023a, January 8). *What is CRISP DM?* Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/

5. Marsja, E. (2021a, April 15). *How to Create Dummy Variables in R (with Examples)*. Erik Marsja. https://www.marsja.se/create-dummy-variables-in-r/