# Introduction

The goal of this visualisation is to help the viewer to compare different aspects such as salary, education, technical language and decide on whether to choose Data Science as a career and in that sense, which country is better to pursue Higher Education. Back in 2012, Harvard Business Review declared: “Data Scientist: The Sexiest Job of the 21st Century.” Authors Thomas Davenport and D.J. Patil (now the U.S. Chief Data Scientist at The White House) devoted much of their article to defining what a data scientist does. For data scientists, it is exciting to see salaries, job opportunities, and job satisfaction on the rise and all the visualisations in this report will help the viewers to reach to a conclusion about choosing Data Science as a career.

## Data Description

The basic dataset is taken from the Stack overflow Annual Developers Survey (2017-2020), The survey is done every year and for the purpose of visualisation we will only consider the respondents that considered themselves already in a data related job (Data Scientist, Machine Learning Specialist, Database Administrator, Data Analyst, Business Analyst and Data Engineer). In the survey different questions are asked, such as “Annual Salary?”, “Years of experience in Coding?” etc.

The Dataset can be downloaded from the following URL: <https://www.kaggle.com/phuchuynguyen/datarelated-developers-survey-by-stack-overflow>

The original dataset contains 303592 rows of data. The rows and columns of the dataset has been updated based on the feedback received for the proposal.

The data dictionary below shows the key original variables

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Data Type** |
| ConvertedComp | Salary converted to annual USD salaries | Integer |
| Country | Country of Job | Nominal |
| DatabaseDesireNextYear | Names of databases planning to work over in next year | Nominal |
| DatabaseWorkedWith | Databases worked on in current and last year | Nominal |
| DB | Number databases worked on | Integer |
| DevType | Roles worked | Nominal |
| EdLevel | Level of education | Nominal |
| Employment | Type of Employment | Nominal |
| Hobbyist | Is developer coding as hobby | Nominal |
| JobSat | Level of job satisfaction | Nominal |
| LanguageDesireNextYear | Programming Languages planning to work on in next year | Nominal |
| LanguageWorkedWith | Programming Languages worked on in current or past year | Nominal |
| OrgSize | Size of Organisation | Nominal |
| ProgrammingLanguages | No of programming languages worked on | Integer |
| UndergradMajor | Major for Undergrad | Nominal |
| Year | Year of survey | Date |
| YearsCodePro | Years of coding experience as professional | Integer |

Some of the columns in the original dataset are multi-valued, to get more value out of them, top 3 values from those columns have been extracted and converted into Boolean columns. After analysing the data using python, only those countries which have at least 50+ survey has been kept, all countries having less than 50 surveys are removed to keep the analysis free of bias. The additional columns are described as follows:

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Data Type** |
| DB | No of Databases worked on in current and last year | Integer |
| ProgrammingLanguages | No of Programming Languages worked on in current or past year | Integer |
| MSSQL | Has worked on MSSQL? | Boolean |
| MongoDB | Has worked on MongoDB? | Boolean |
| MySQL | Has worked on MySQL? | Boolean |
| Python | Has experience in Python? | Boolean |
| R | Has experience in R? | Boolean |
| Scala | Has Experience in Scala? | Boolean |

## Persona and Questions

The viewer who is a student from a non-technical background, wishes to understand and plan their higher studies abroad and wish to go for Data Science as a career. The viewer has two specific questions:

1. What is the average salary for a Data Scientist and factors affecting the salary?
2. Which are the 5 best countries for higher education in Data Science based on salary and job satisfaction?

N:B – The questions have been reiterated and updated according to the feedback received for the proposal.

### Software and Hardware Requirements

There are some basic software and hardware requirements to create and run this project on a laptop. The author has used Tableau Desktop; this can be downloaded from Tableau official website.

To run Tableau Desktop on a desktop/laptop these are the basic required specifications:

#### Windows:

* Microsoft Windows 7 or newer (x64)
* 2 GB memory
* 1.5 GB minimum free disk space
* CPUs must support SSE4.2 and POPCNT instruction sets

#### Mac:

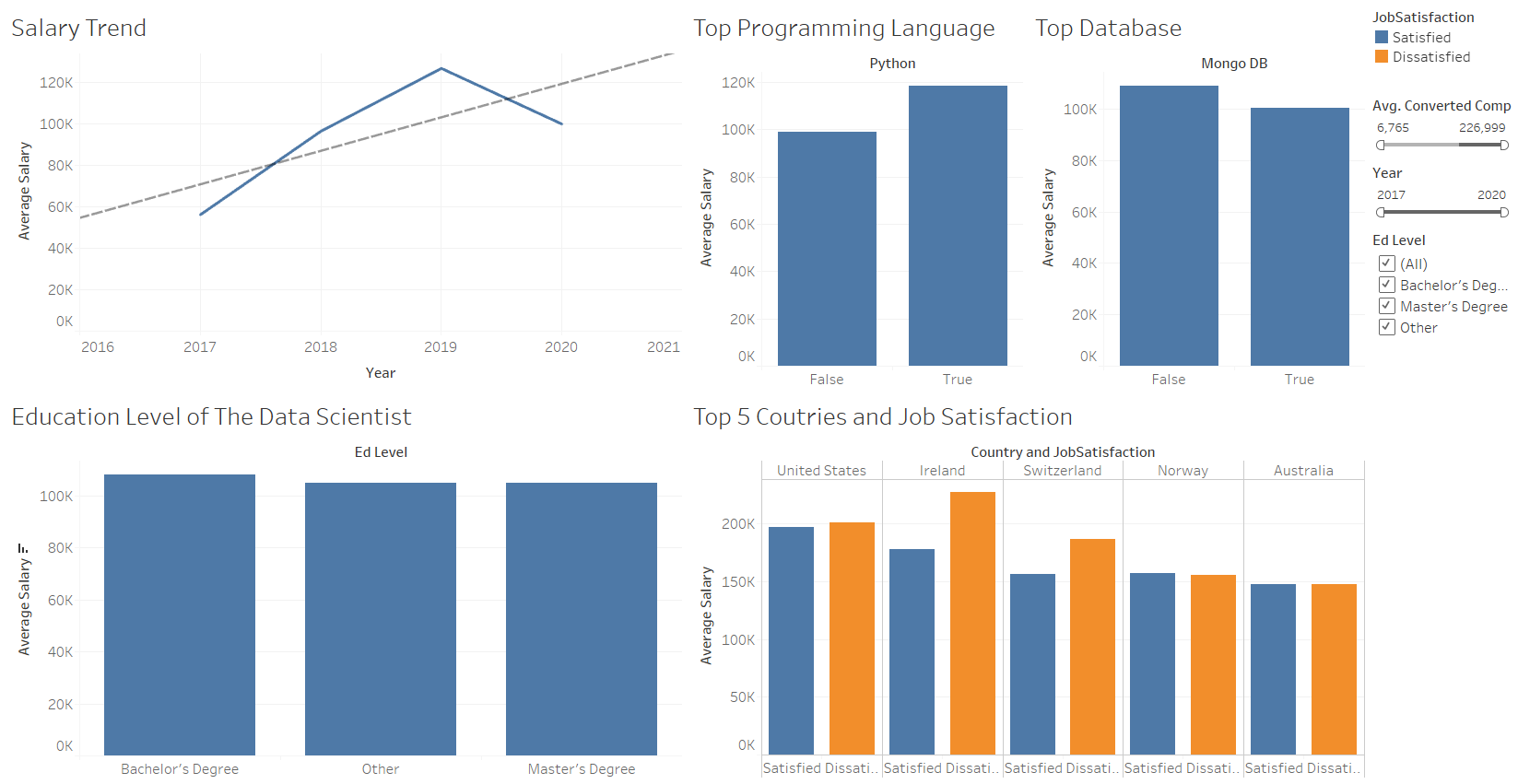
* macOS High Sierra 10.13, macOS Mojave 10.14 and macOS Catalina 10.15
* Intel processors
* 1.5 GB minimum free disk space
* CPUs must support SSE4.2 and POPCNT instruction sets

# Design

2 Difference in pay for Python and Mongo DB experts and non-experts.

4 Multiple interactive filters.

1 Avg. Salary and trend over the year



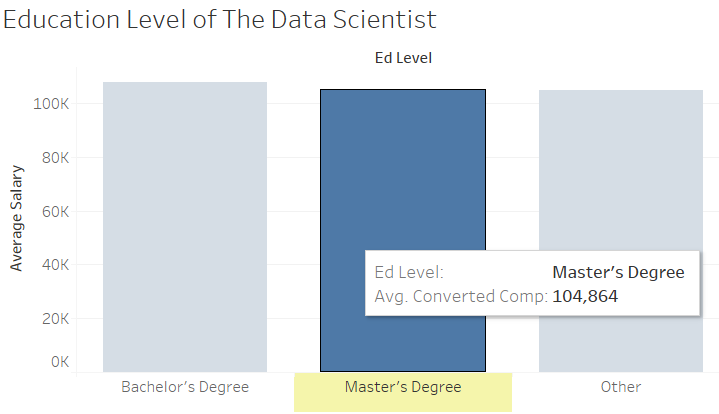
5 Job satisfaction for Top 5 countries by average salary.

3 Education level and Avgerage Salary

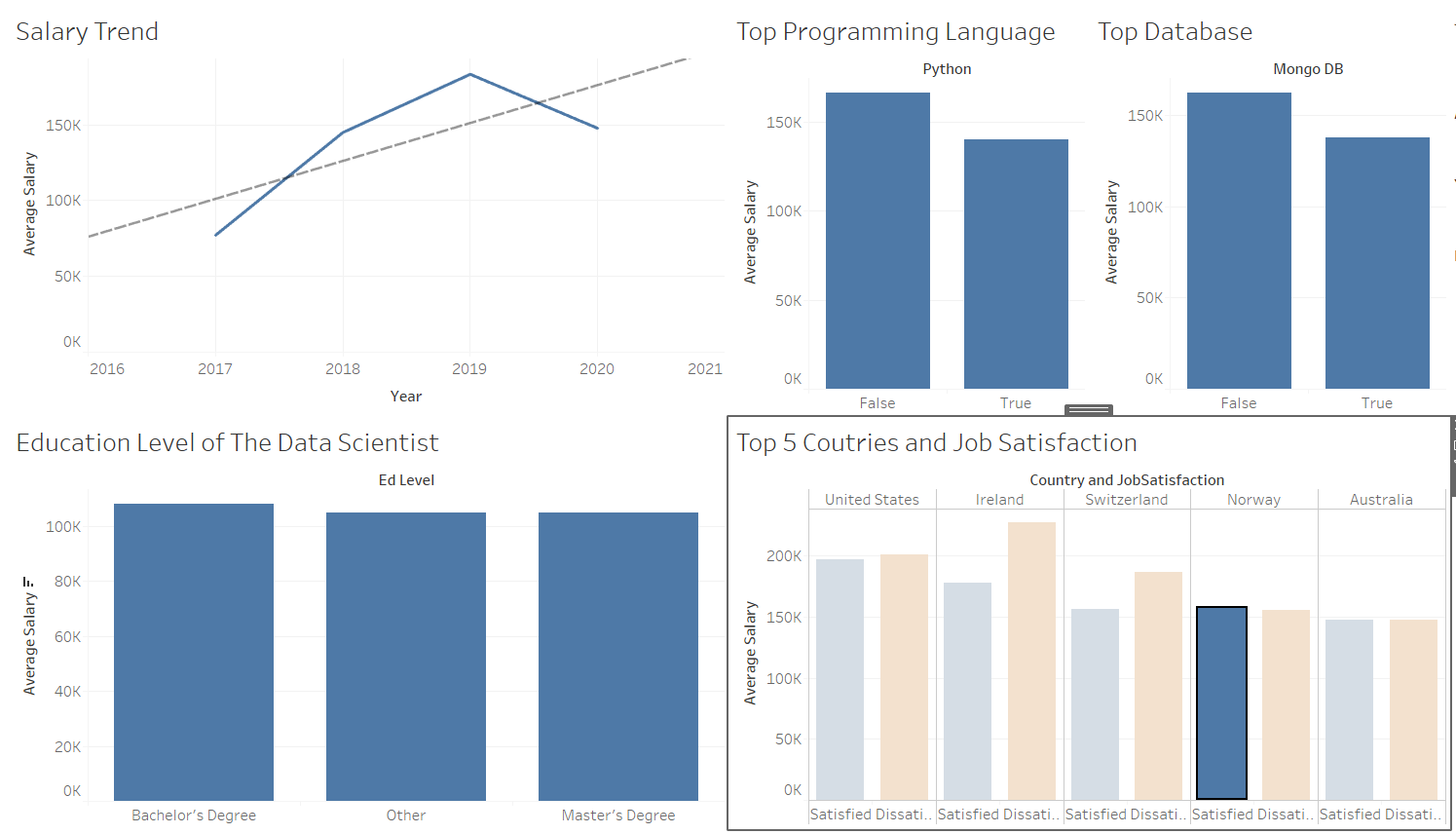
The final dashboard provides a lot of interactivity, if the viewer clicks on any of the view, then values are

shown specific to that value. If the viewer hoovers over a specific region, then detailed values related to that

region is highlighted.



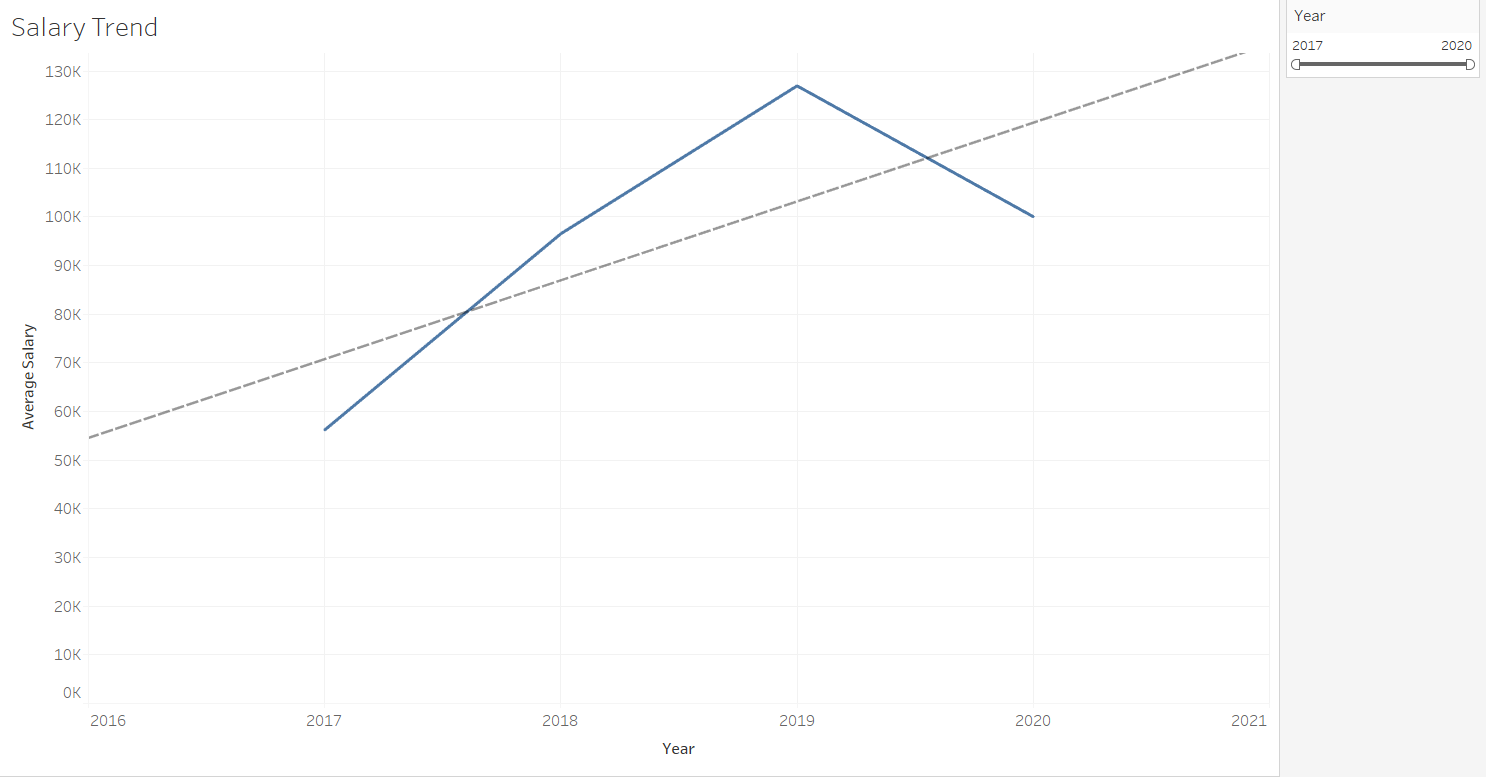
The view above shows the average salary of a data scientist with a bachelor’s degree, the value is highlighted then the mouse is hovered over the bar. The below shown view is displayed when the viewer clicks on the satisfied data scientists in Norway and all other views also interactively get updated and show data only related to satisfied data scientists in Norway.



# Implementation

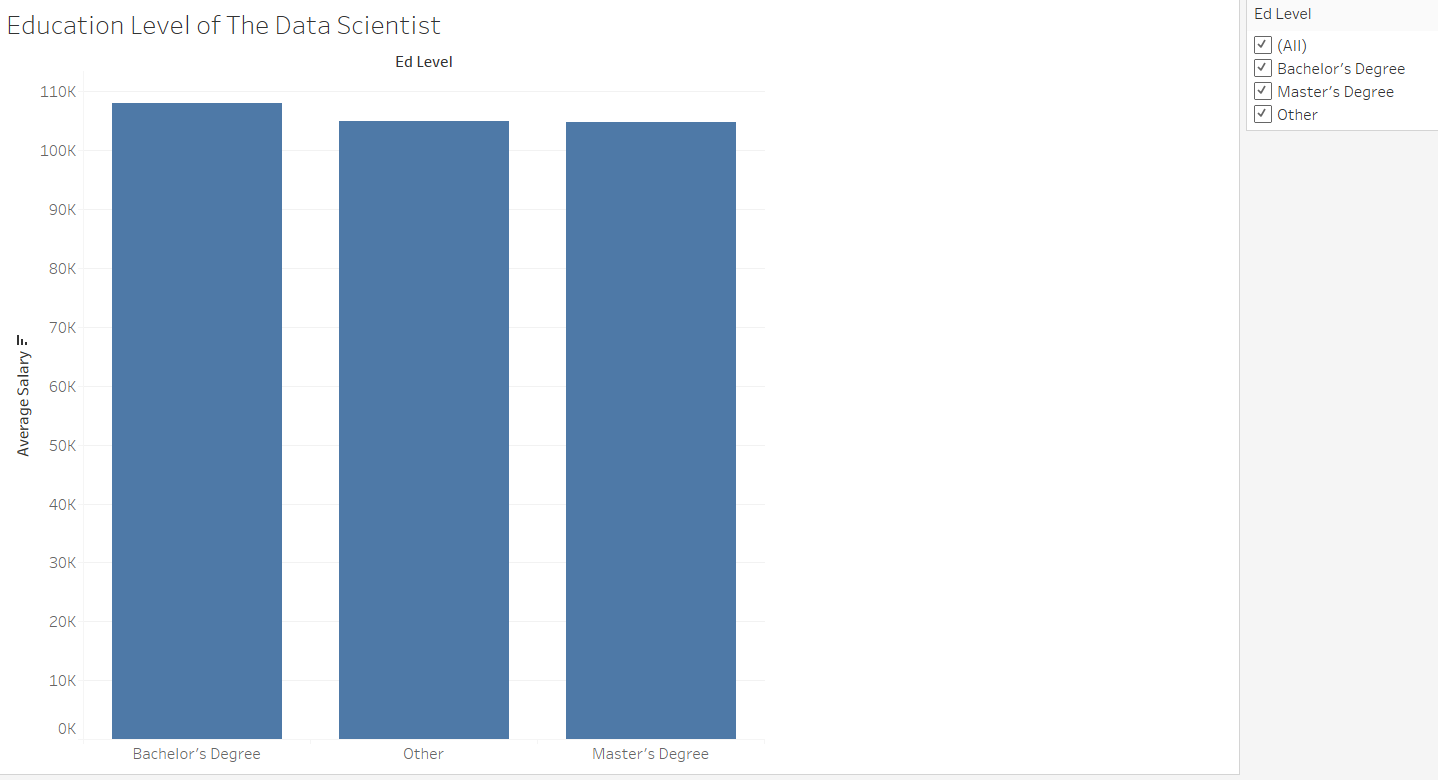
The final dashboard is designed to answer specific questions related to Data Science as a career and in order to reach at to those answers, author has created 5 different views including different aspects such as lineplot of average salary over the years, bar chart of average salary with different education level, bar chart for top five countries with job satisfaction level etc., The implementation of all the views is as follows:

### Salary Trend View



This view aims to display the annual average salary of a data scientist all around the world from year 2017 to year 2020. A statistical technique known as linear regression is also applied to display the trend of salary over the years and this dotted line also tried to predict the salary for next years based on data available for previous years. To create this view, average of “Converted Comp” is taken as row field and “Year” has been selected as the column fields. A filter has been applied on the “Year” field and it can be used to select and display data for a specific year range. To add the trend line Analytics section of Tableau is used and from the Model section a Linear Trend line is chosen.

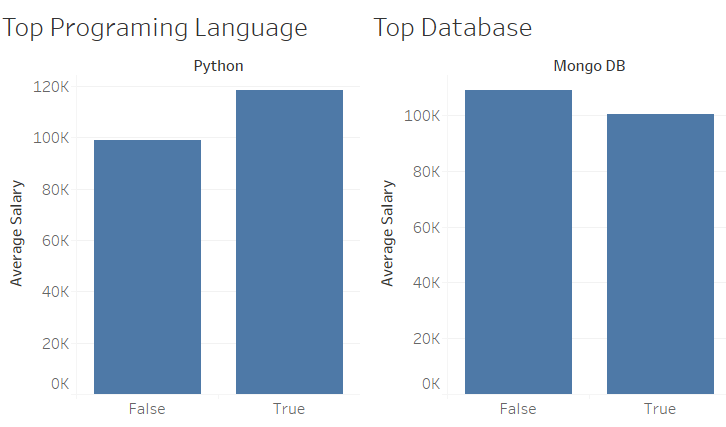
### Education Level View



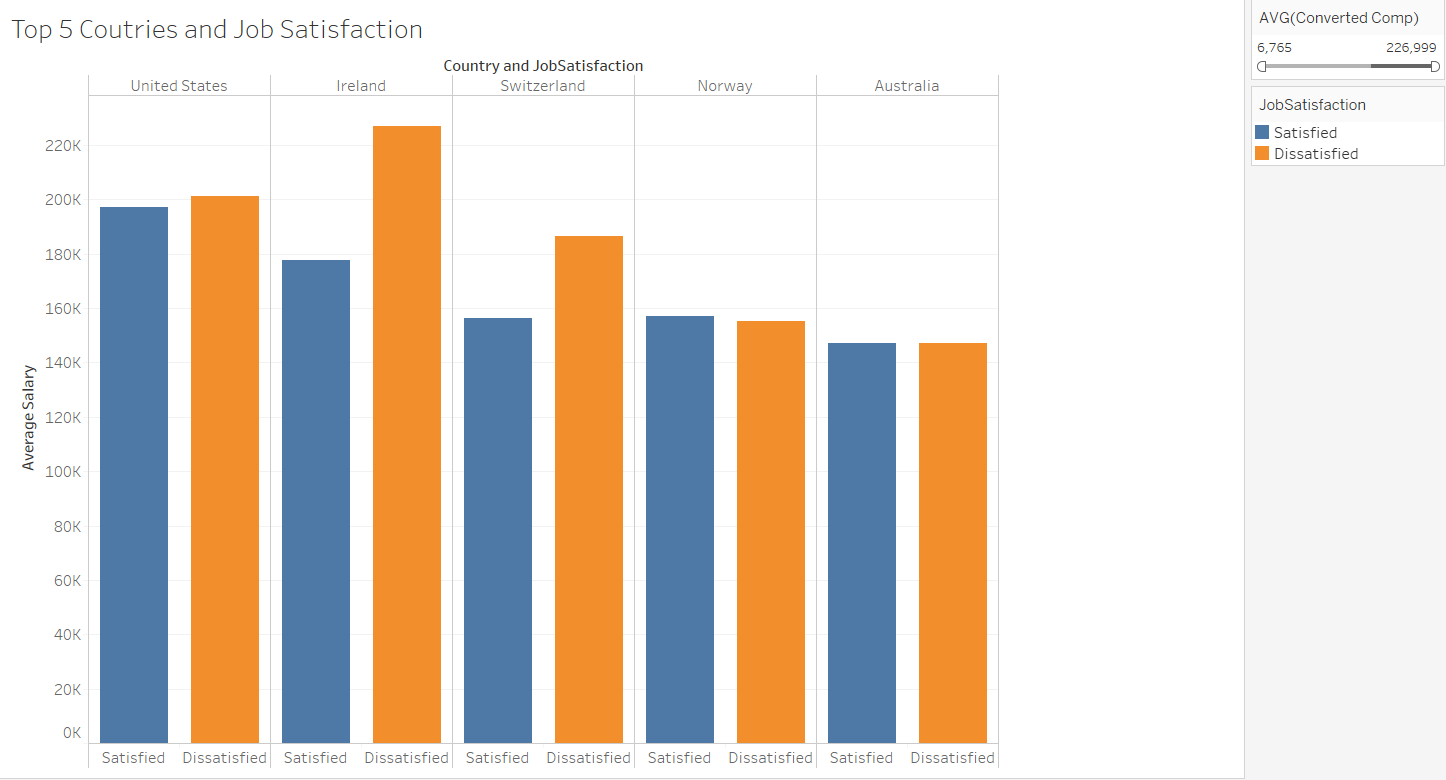
This visualisation is created with an aim to compare the pay scale of different data scientists based on their level of education. There were multiple level of education in the “Ed Level” dimension, all those have been converted into Others and Master’s and Bachelor’s degree has been kept as is. Average of “Converted Comp” is taken as row field and “Ed Level” has been selected as the column fields, the bars are in descending order of average salary. A filter has been applied on the categorical variable “Ed Level” to select data for a specific level of education.

### Top Programming Language and Top Database View

Two views below show the average salary of a professional who has working experience in the top programming language “Python” and top database “Mongo DB”. These views simply select average of “Converted Comp” as row value and boolean dimensions “Python” and boolean dimension “Mongo DB” as columns value, respectively. There is no additional filter have been added for these views. Initially the dimensions “LanguageWorkedWith” and “DatabaseWorkedWith” had multiple values and it was quite difficult to extract any relevant information from these. The feedback suggested to extract top-k values from these and the author used Python programming to extract top-3 values from both “LanguageWorkedWith” and DatabaseWorkedWith” converted them into boolean values and the view Top Programming Language and Top Database View helps in comparing salary for an expert and non-expert of top technical skills.



### Top 5 Countries and Job Satisfaction



A relationship between average annual salary and job satisfaction is presented in this view. Average of “Converted Comp” measure is selected as row value and there are 2 values selected for columns for this view, dimension “Country” and dimension “JobSatisfaction”. “Job Sat” is a multi-values dimension which has multiple level of job satisfaction, a calculated field “JobSatisfaction” has been derived from “Job Sat” and has only two values “Satisfied” and “Dissatisfied”.

The applied formula is:

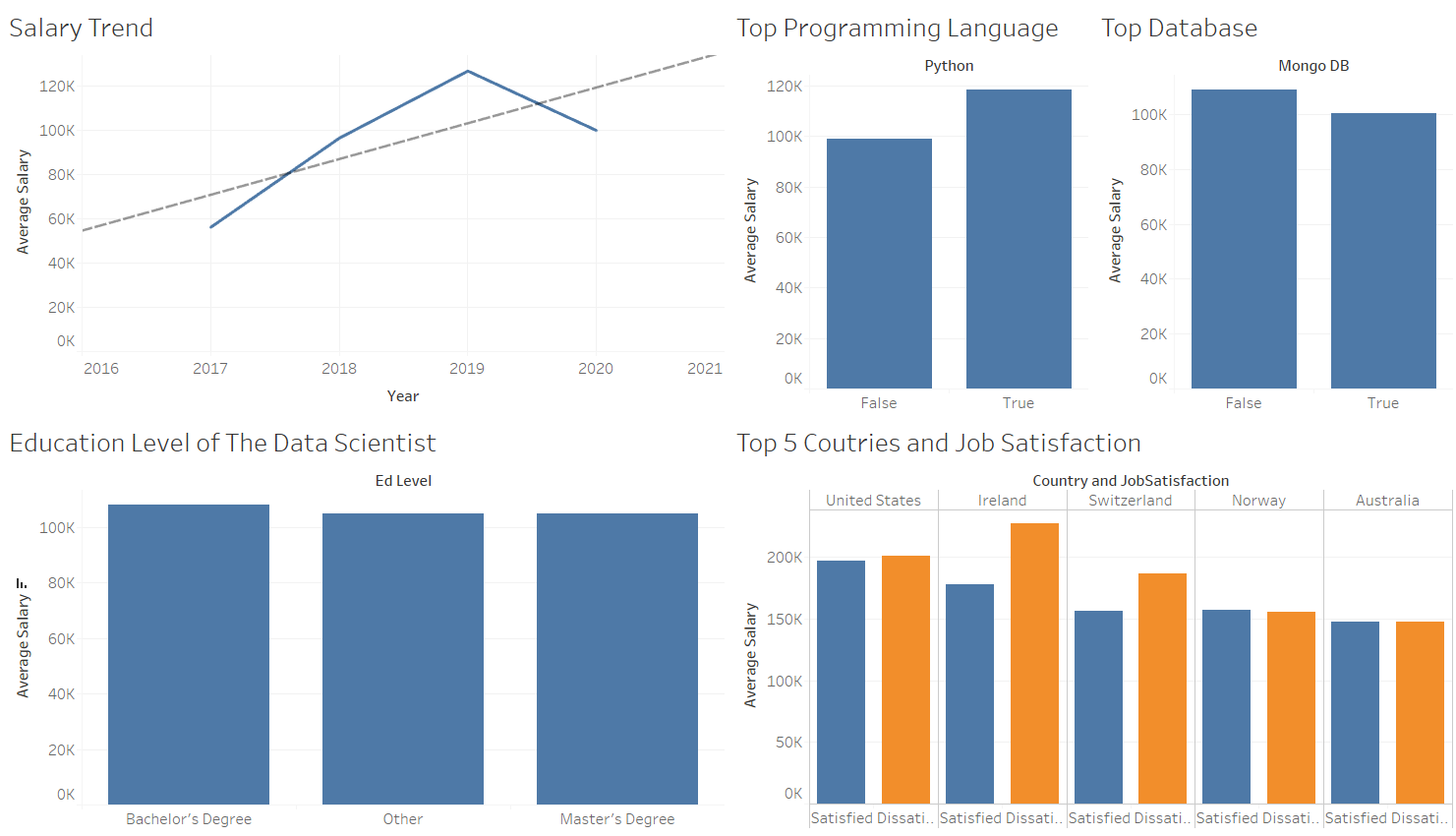
IF CONTAINS([Job Sat],"Neither" ) THEN "Satisfied"

ELSEIF CONTAINS([Job Sat],"dissatisfied" ) THEN "Dissatisfied"

ELSE "Satisfied" END

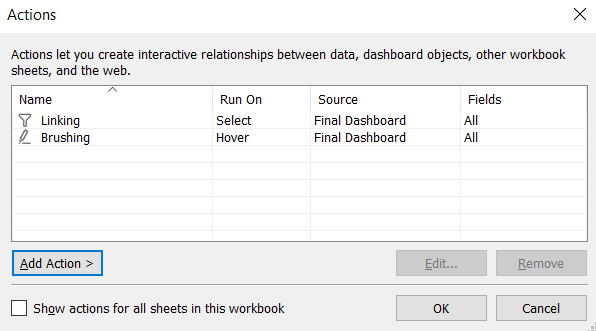
A colour hue has been applied in marks section on “JobSatisfaction”, Orange displays Dissatisfied data scientists and Blue displays satisfied data scientists. A filter on average “Converted Comp” is applied so that the viewer can select a specific range of salary.

### Final Dashboard



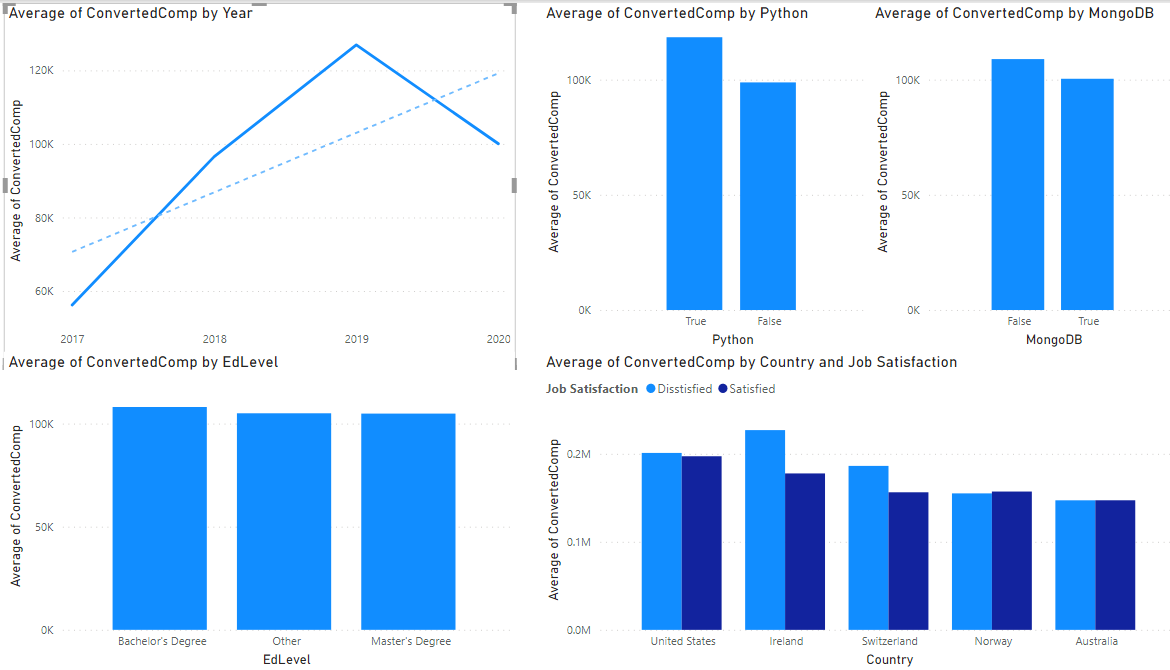
The final view shows the relationship between different aspects such as salary, education, and job satisfaction in relation to salary of a data scientist. Brushing and linking is an impressive filtering technique in interactive data visualisation. In this technique a viewer can filter data by interacting directly with marks in the views, rather than using separate filter widgets. Brushing means an action to select one or more data points in a view. Linking is the concept of coordinated views which means that changes to the configuration of one view (e.g. selection of data points) are reflected equivalently in other views. User can hoover over any data and the dashboard will show detailed values for that data. Viewer can click/select on any of the bar or data point and all other views will get changed accordingly. To apply this functionality in user can follow these steps:

Dashboard -> Actions -> Add Action -> Then select required functionality accordingly.



The author has applied two actions to this dashboard for interactivity.

### Power BI Implementation



Power BI is another tool by Microsoft which can be used to visualize data. Here, the user has implemented the same views using Power BI. For view1, Select “Year” as axis and average of “Converted Comp” as value, from analytics section choose a trend line to draw a linear regression line. For view2, select “Ed Level” as axis and average of “Converted Comp” as value. For view3 and view4, select average of “Converted Comp” as value and “Python” and “MongoDB” as axis. For view5, create a new column based on following formula:

“Job Satisfaction = IF(CONTAINSSTRING(survey\_updated\_FinalData\_2035261[JobSat],"\*neither\*"),"Satisfied",IF(CONTAINSSTRING(survey\_updated\_FinalData\_2035261[JobSat],"\*dis\*"),"Disstisfied","Satisfied"))”

and select average of “Converted Comp” as value and “Country” as axis with a filter Top 5 values based on average “Converted Comp” and “Job Satisfaction” as legend.

# Evaluation

Goal of an infographic is to represent information, data, or knowledge in order to communicate complex information quickly and clearly. The main task for this infographic is to provide some knowledge about average salary for a Data Scientist from data taken from a around the world and look at factors that affect the salary. Additionally, fetch top 5 countries for a data scientist and the job satisfaction in those countries.

The views created in the report elegantly provides information about the average salary for the job role data scientist for past four years, initially there was a rise in the salary but with time the average salary value started to dip down a bit. There is a significant difference in the salary of a Python expert data scientist as compared to a non-expert, although having a database as a skill is a plus but it does not make any significant difference on the average salary. Education level seems to have very minute significance on the average salary, but surprisingly professionals with a master’s degree earn less than those who have a bachelor’s or some other degree.

Top 5 countries with best average salary are United States, Ireland, Switzerland, Norway, and Australia, United States being the best and Australia being the fifth best country. For most of these countries the average salary of dissatisfied people is more than satisfied people. Shockingly, in Ireland more people are dissatisfied with their jobs even after having quite a high pay.

### Strength

The project successfully explains the required information and it can give a good idea about the aspects of data science as a career. The viewer who is a student can look at the different views and decide about their higher studies abroad. The viewer can narrow down the options to top 5 countries and based on salary and job satisfaction, decide on one country.

### Weakness

The data for most of the columns is nominal, which makes it hard to make different comparisons. Although,

enough integer columns are also available. Number of surveys from different countries is also a drawback. Countries with less than 50 surveys have been removed from the data but still the number of surveys are quite varied in different countries.

# Reflective Discussion

Overall, the project is quite informative and provides a deep insight into a data scientist’s career andpay scale. Tableau is a very easy to use software and I found it quite impressive in many ways. User can easily add multiple data sources for a project. Tableau provides a great range of graphical representation options. Tableau community provides plethora of tutorial videos which can be viewed to learn to apply different tasks in Tableau. User can easily do multiple calculations like average, sum, count, median etc. Tableau allows user to apply some analytical functions such as trends, clustering, and quantiles.

Although, there are numerous advantages of using Tableau, there are a few drawbacks. Tableau is little difficult to integrate with code environments. I used python on the original data to apply some cleaning. Tableau desktop is a heavy software and used a lot of memory which makes the desktop slow. Creating calculated fields is a great function but if the user is not so familiar with the coding, then it makes it a bit difficult to create complex calculated fields.

Tableau Server is a product from tableau which I want to use in the coming future. Tableau Server helps the user to publish dashboard, worksheets. Once the workbook is uploaded on the server, the recipient need not have Tableau Server installed in their PC. All that they need is Login credentials. Tableau Admin can set permission for each user, whether they can view, make changes, edit etc. This saves memory and space for the user.

# References

Davenport, T.H. and Patil, D.J.(2012). Data Scientist: The Sexiest Job of the 21st Century. [online] Available at: <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century> [Accessed 18 April 2021].

Laurent, B. (2020) Data Scientist: Still the “Sexiest Job of the 21st Century? [online] Available at: <https://www.dominodatalab.com/blog/data-scientist-still-the-sexiest-job-of-the-21st-century/> [Accessed 19 April 2021].

Lab Notes [online] Available at: <https://blackboard.brunel.ac.uk/> [Accessed 19 April 2021].

Microsoft Power BI (2020). [online] Available at: <https://docs.microsoft.com/en-us/power-bi/transform-model/desktop-analytics-pane> [Accessed 22 April 2021].

Microsoft Power BI CONTAINSSTRING (2020). [online] Available at: <https://docs.microsoft.com/en-us/dax/containsstring-function-dax> [Accessed 22 April 2021].

Microsoft Power BI Create Calculated column (2019). [online] Available at: <https://docs.microsoft.com/en-us/power-bi/transform-model/desktop-tutorial-create-calculated-columns> [Accessed 22 April 2021].

Microsoft Power BI Community (2017). [online] Available at: <https://community.powerbi.com/t5/Desktop/if-then-else/td-p/117999> [Accessed 22 April 2021].

Phuc, NH. Kaggle (2020). [online] Available at: <https://www.kaggle.com/phuchuynguyen/datarelated-developers-survey-by-stack-overflow> [Accessed 3 March 2021].

Slaughter, A. (2016).Using Brushing and Linking in Tableau. [online] Available at: <http://adrienneslaughter.com/2016/01/27/using-brushing-and-linking-to-understand-prediction-performance/> [Accessed 19 April 2021].

Tableau Products. (2020). [online] Available at: <https://www.tableau.com/en-gb/products> [Accessed 19 April 2021].

Tableau TechSpecs. (2020). [online] Available at: <https://www.tableau.com/en-gb/products/techspecs> [Accessed 20 April 2021].

Tableau Desktop. (2020). [online] Available at: <https://help.tableau.com/current/pro/desktop/en-us/functions_functions_logical.htm> [Accessed 18 April 2021].