Tableau – Salary data set

Antonio Guadagno – 10561018 – antonio.guadagno@mail.polimi.it

Antonio Guadagno – 10561018 – antonio.guadagno@mail.polimi.it

Salary data set description

When a US company wants to hire someone from outside of the United States for a technical position, they have to file an application to the United States government to get a green card or visa for the foreign applicant. These applications allow the US government to track who is entering and leaving the country for work-related reasons and ensure that immigrants are neither being taken advantage of nor causing adverse effects for U.S. workers. To ensure equity for US and non-US workers, companies have to state how much they are planning on paying the employee every time they submit a visa or green card application. They also have to state the average amount an employee with similar skills and background typically gets paid for the same position, a figure called "the prevailing wage." This publicly available data provides a unique view into what types of salaries you might encounter for different data–related jobs in the US.

Source: the original data was compiled by the US Department of Labor's Office of Foreign Labor Certification (http://www.foreignlaborcert.doleta.gov/performancedata.cfm

1 – Change data types (if needed)

Looking at the data set, we can notice some wrong data types. We can change data types directly from Tableau.

- **Case received date**: string → date
- **Decision date**: string → date
- **Prevailing wage submitted**: string → number (decimal)
- **Experience required**: string → number (whole)
- **Paid wage per year**: dimension (discrete) → measure (continuous)
- **Prevailing wage per year**: dimension (discrete) → measure (continuous)

- **Dependent variable**: paid wage per year (rows)
- **Independent variable**: job title subgroup (columns)

Median wage per subgroup

Job Title Subgroup
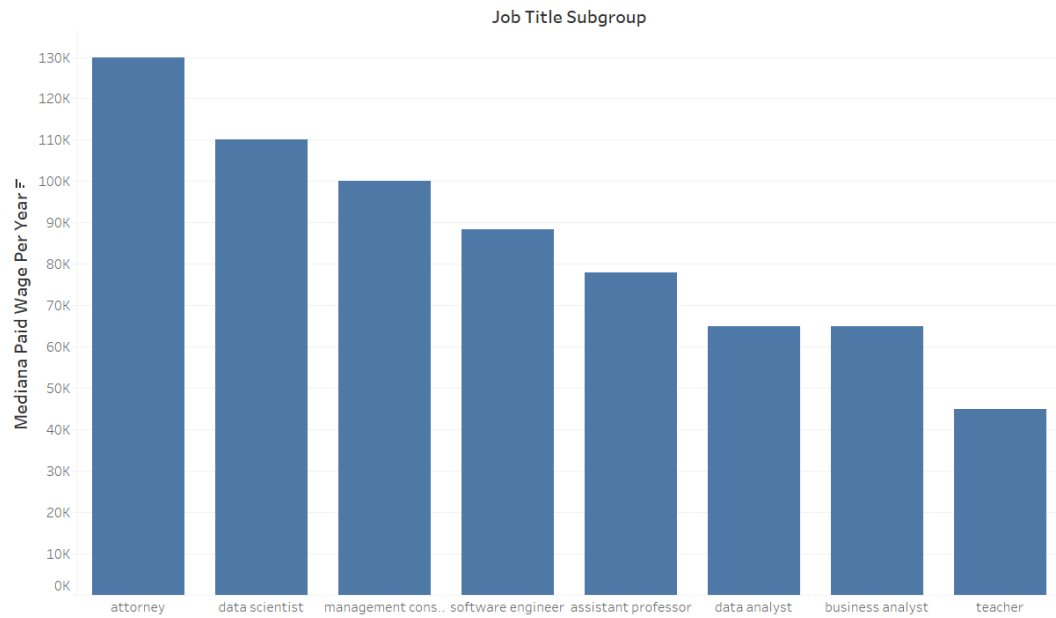
3 – Make a rational hypothesis an verify it (can be correct or incorrect)

Knowing that green cards cost more than visas, I can make the following hypothesis:

"A company will pay for a green card only to hire extremely competent people, therefor wages of people having a green card will be higher with respect to wages of people having visas."

To verify my hypothesis, I plot the median paid wage for job title subgroup and visa class.


Median wage per subgroup and visa
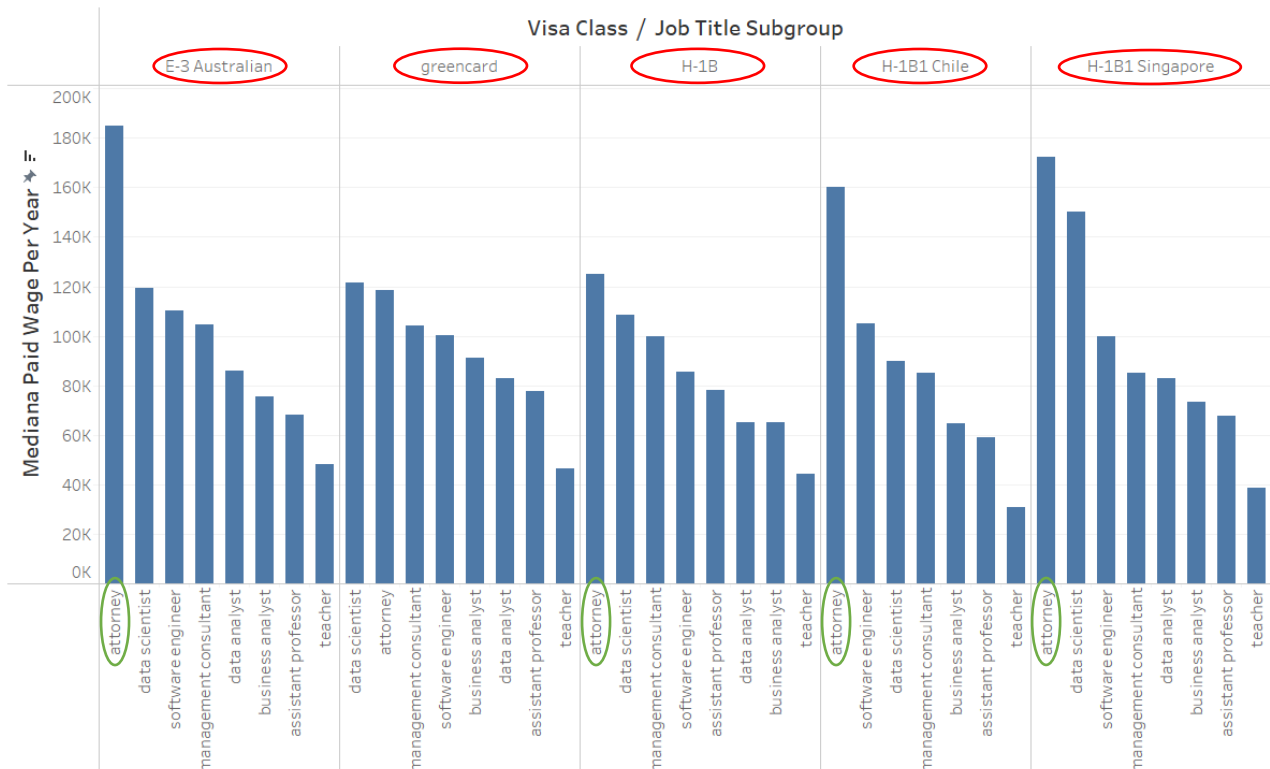
Looking at the plot, I can say that my hypothesis was not correct: in most of the cases, attorneys are the ones who earn the most, regardless of their visa class.

## 4 – Are medians reliable? How do you check medians' reliability? What else can you say?

To understand if medians are reliable, we can plot and analyze the **standard deviation**. The higher the standard deviation is, the less the median is reliable.
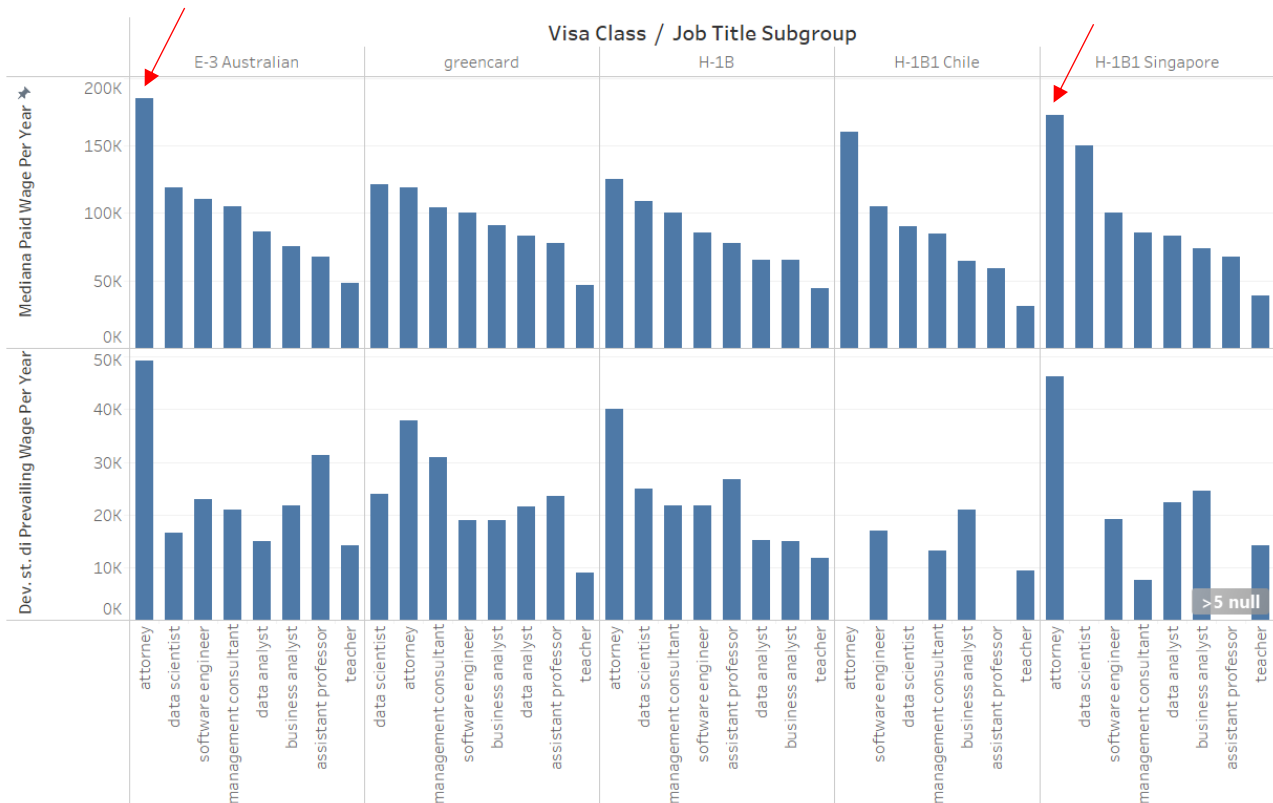
From this plot we can notice that the highest medians are associated to the highest standard deviations, therefor those medians are not reliable. High standard deviations suggest us that **outliers** may be present.

Although a bubble plot is not the best choice for this kind of query, here is the plot we obtain. Please notice that, to improve the chart readability, I have associated each job title subgroup with a different color. The bigger the bubble, the higher the median paid wage.
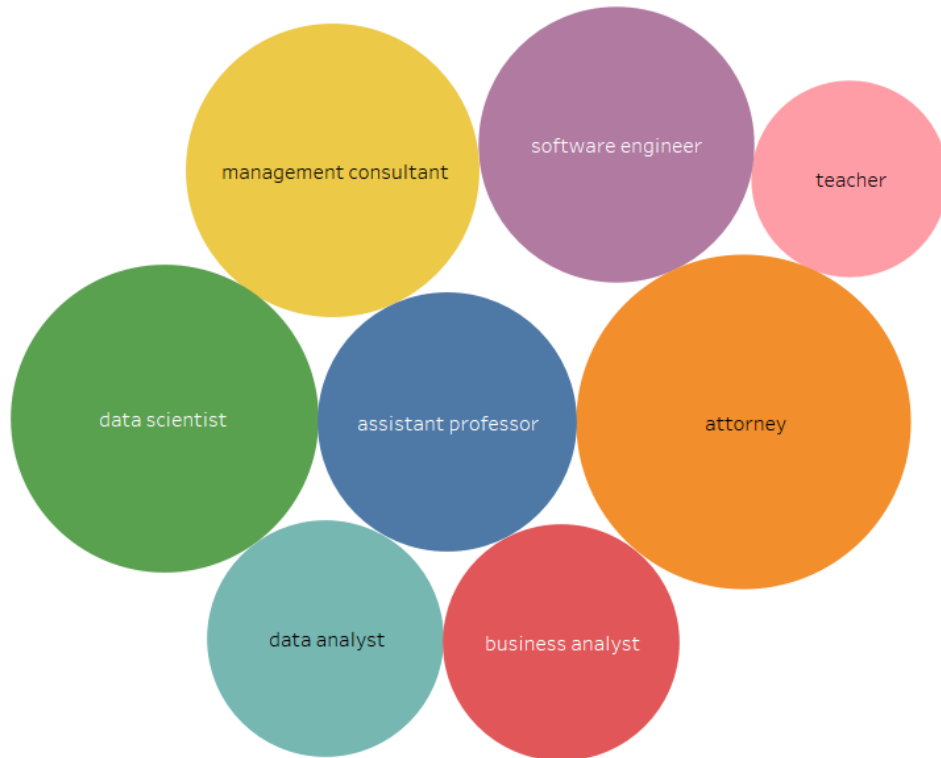
6 – Identify outliers (if present) and remove them

As we have already said, high standard deviations suggest us that outliers are present. The highest standard deviations are associated to attorneys' wages, so I expect to find outliers among them.

The best way to identify outliers is using **scatter plots** (visual inspection).
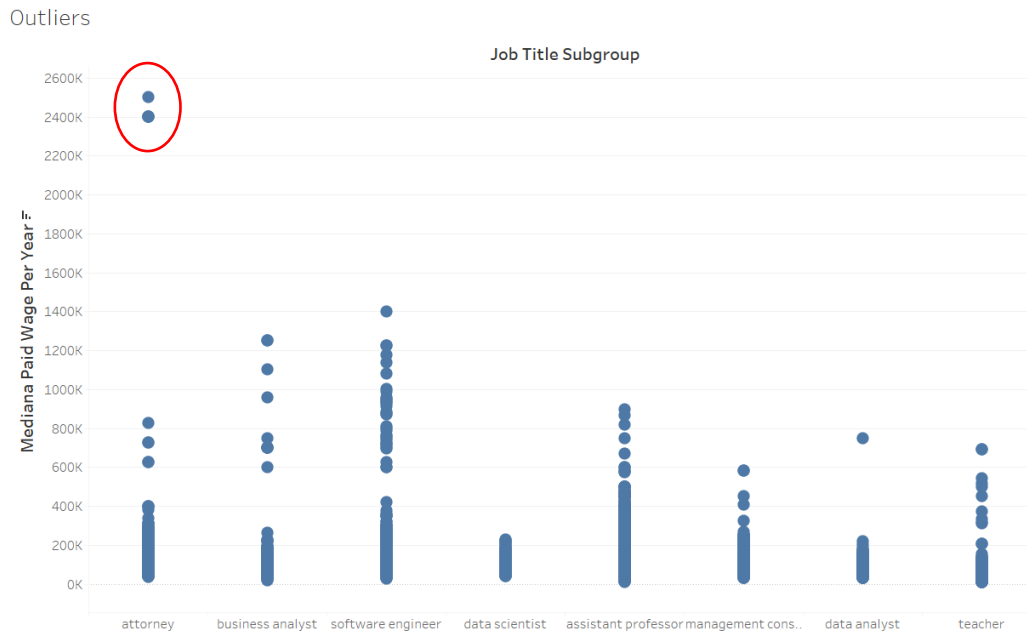
To remove outliers: add "case number" to "details", select the outliers, group by "case number", drag-and-drop "case number (group)" into filters, right click on "case number (group)", select "show quick filter", select "other" only. In this way, outliers won't be visible.
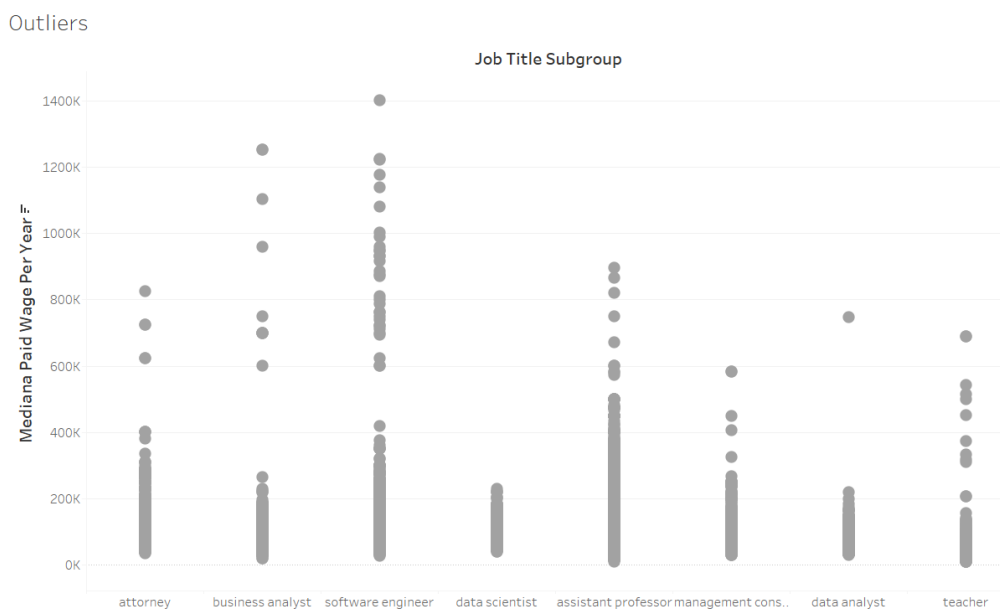
7 – Your goal is to understand if, given a certain job, the wage associated to it changes from a country to another. Please consider only the following countries: California, Washington, North Carolina, Colorado, Texas, New York, Massachusetts, Alabama, Maine.

We need to apply a filter on "Work state". The problem is that, in the data set, sometimes they use the full name of a country (ex: "California") and sometimes an abbreviation (ex: "CA"). To solve the problem:

Work state, create, group, create groups (and rename them) for those countries we're interested in.

In this way we obtain "work state (group)". We drag-and-drop "word state group" in "filter" selecting only the states we're interested in.

Note: the free version of Tableau doesn't allow exports, therefor I'm able to show only a part of the obtained plot, otherwise it won't be readable.
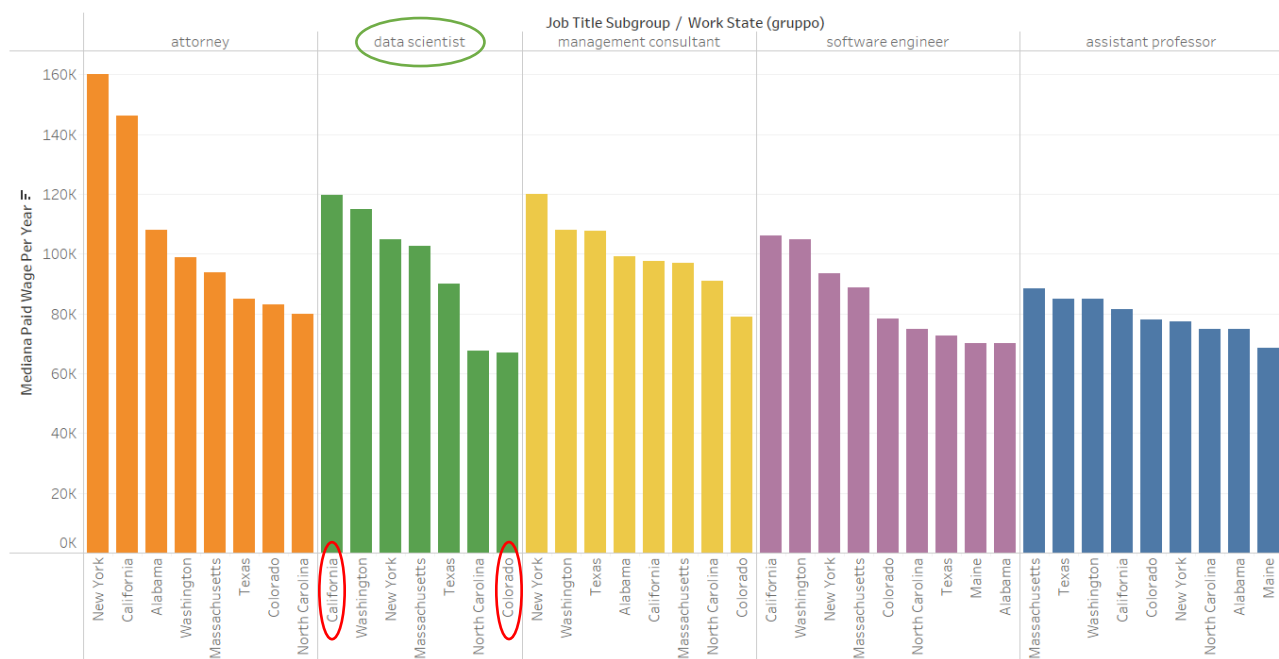
Looking at the plot we can say, for instance, that for a data scientist is a good idea to move from Colorado to California.

8 – Does the wage per year increase proportionally with the years of experience?

To answer to the question, we will plot the median paid wage for year in function of the case received data. So, "case received date" basically is the independent variable (column) and the median of "paid wage per year" basically is the dependent variable (row).

"Case received date" is a date and therefor is treated as a dimension (discrete variable). If we plot the median paid wage for year, we will obtain a disconnected line.

To solve the problem (and so, to obtain a continuous like), we need to treat the "Case received date" as a measure (continuous variable). To do that we drag-and-drop "Case received date" in columns, we click on the down arrow, and we select "year". "Case received date" will change its color from blue to green: now it is a measure and lines will be connected.

The plot that we obtain is the following one:

Median paid wage for year in function of case received data
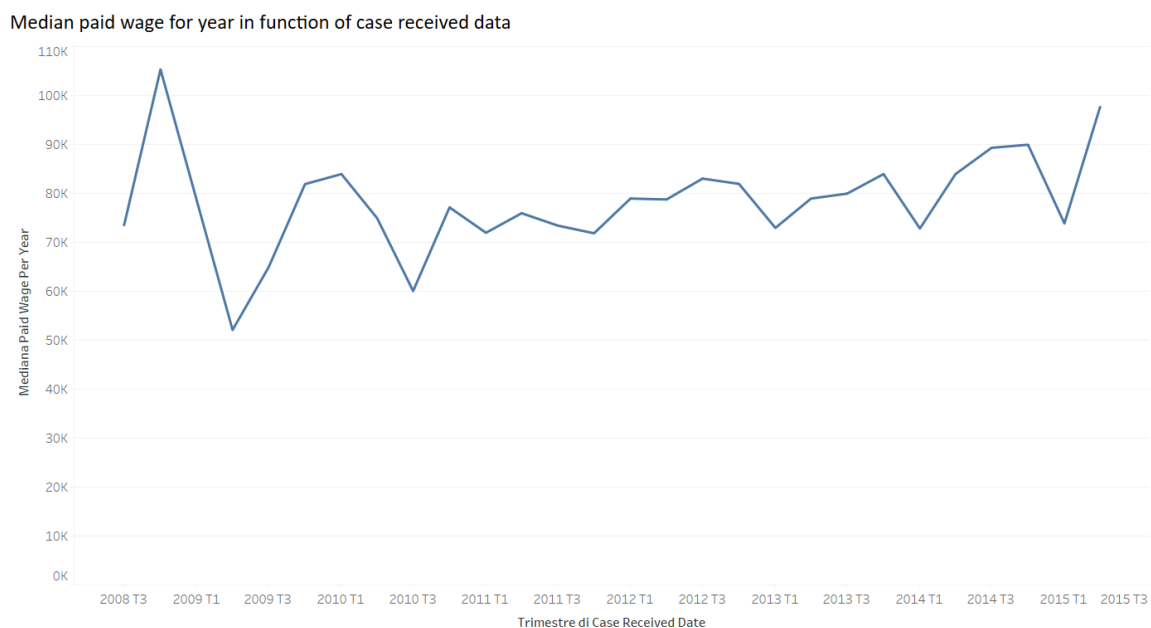
We can say that the median paid wage per year is not significantly changed through time.

1) **Can you say that their minimum wage is decreasing/increasing?**
2) **Can you say that their maximum wage is decreasing/increasing?**
3) **Can you say that their median wage is decreasing/increasing?**

To answer to this question, the first thing to do is to plot the minimum, the maximum and the median "paid wage per year" per "job category". Then, we drag-and-drop "job title per subgroup" on "color" in order to have data for each job title separately. Finally, we introduce tendence lines and we select only that lines that are associated to data analysts. This is the plot we obtain.
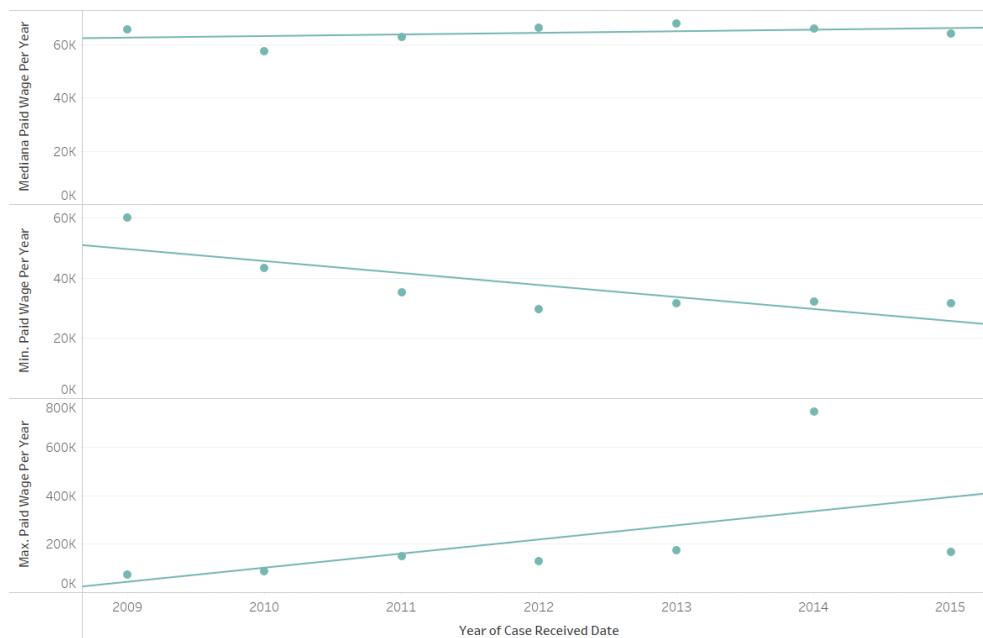
Apparently, the median wage of data analysts is increasing, the minimum wage of data analysts is decreasing, and the maximum wage of data analysts is increasing. Are these hypotheses reliable? We have to check the **p-values**. If we pass with the mouse on the lines, we can read it.

The p-value associated to the Max and the Median graph is higher than 0.05, this means that we cannot be sure of the conclusion we have we came to.

However, the p-value associated to the Min wage per year is 0.03 that is lower than 0.05, so we can say that the minimum wage per year data analysts is decreasing.
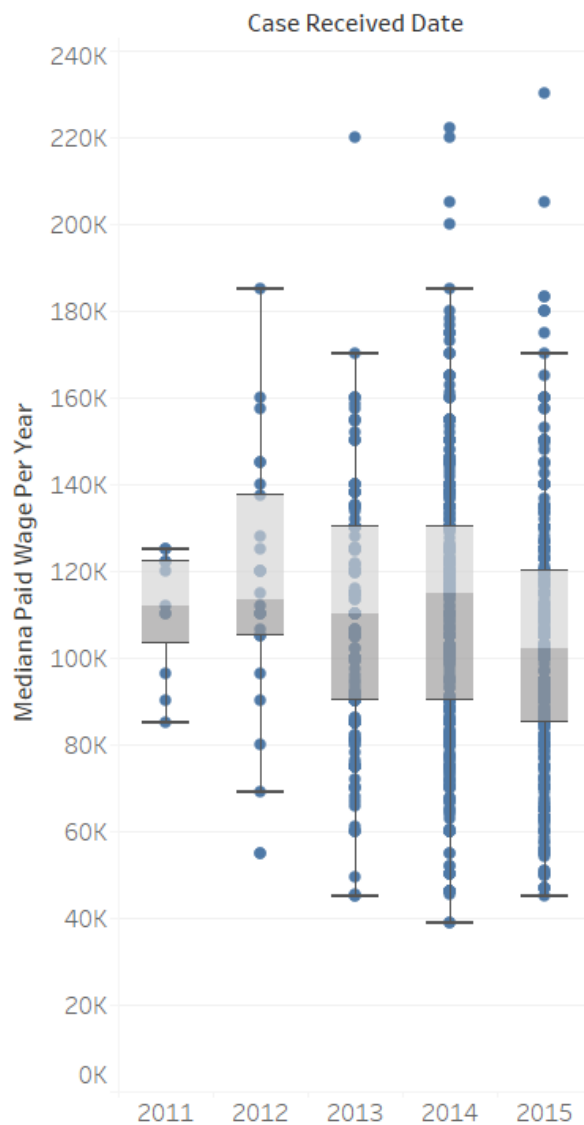
10 – Box plots: median "paid wage per year" for "year" focusing on "data scientists"

In this case we need to consider the **date** as a **discrete variable**.

- Column – Case Received Date
- Row – Median paid wage per year
- Filter – Job Title Subgroup (to select "data scientists").

Then we go to "Analysis" and we click on "Aggregate measures". This is what we obtain:

First, we need to specify that we want to consider Texas and California only. To do that, I've written the following calculation:
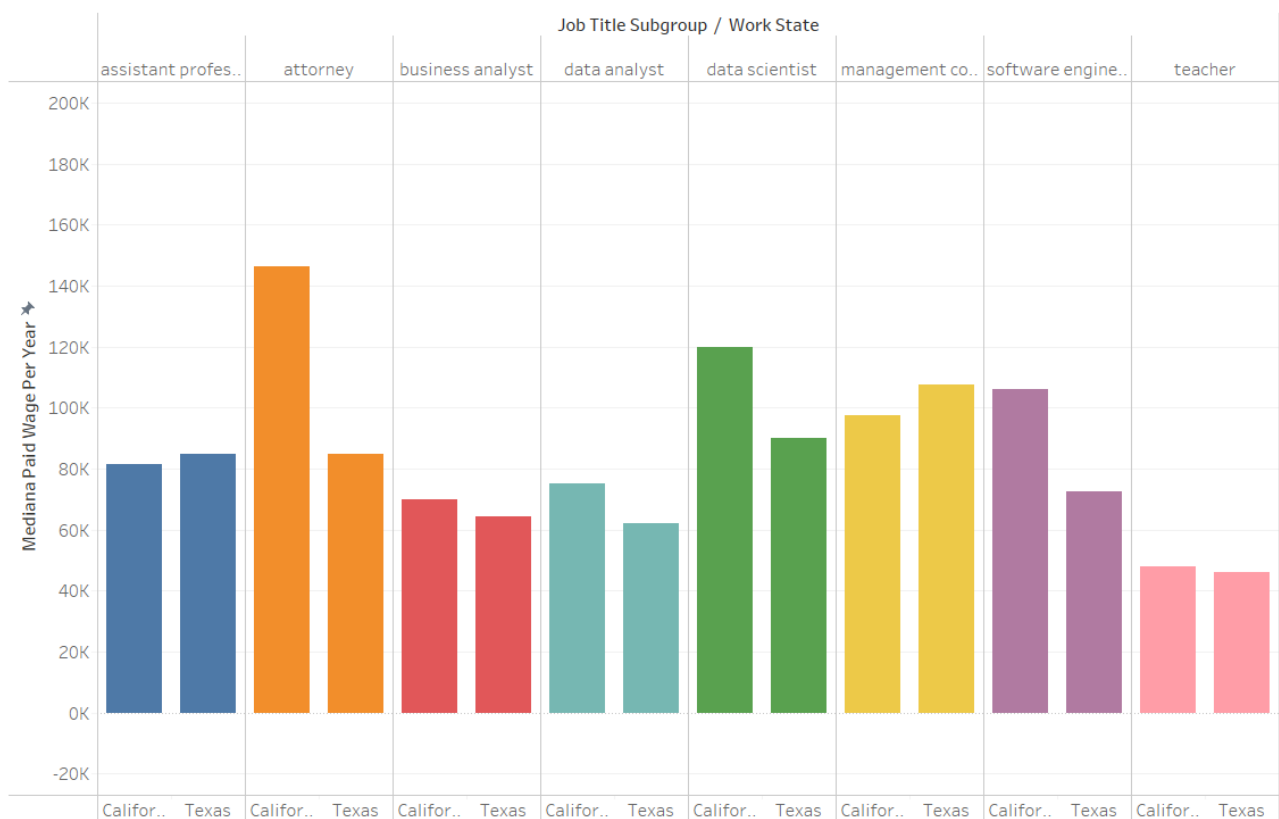
```
States we want

IF ([Work State] = "California"
OR [Work State] = "Texas")
THEN "VIEW"
ELSE "DONT VIEW"
END
```

To obtain the requested plot we have to drag-and-drop "States we want" in "filters" and we have to select "VIEW" only. Then we have to drag-and-drop "Work state" in "columns" and "Job Title Subgroup" in "color". This is the plot we obtain:

Please notice: Tableau Desktop does not allow exports. For this reason, some labels are cut and not completely visible.

First, using calculations, I need to create a variable to compare the prevailing wage per year and the paid wage per year for each employee.

```
Difference between prev

[Paid Wage Per Year]-[Prevailing Wage Per Year]
```

The variable we have just created is called "Difference between prevailing and paid wage".

- If "Difference between prevailing and paid wage" is greater than 0, then the employee is paid more than the prevailing wage per year
- If "Difference between prevailing and paid wage" is lower than 0, then the employee is paid less than the prevailing wage per year

To express this concept, we create a new calculated variable: "Paying less than prevailing wage".

```
Paying less than prevailir

IF [Difference between prevailing and paid wage per year] < 0
THEN "Paid Less"
ELSE "Paid More"
END
```

Now we need to build our plot:

- Columns: job title subgroup
- Rows: Count(Difference between prevailing and paid wage per year)
- Filters: Paying less than prevailing (selecting "Paid less" only)
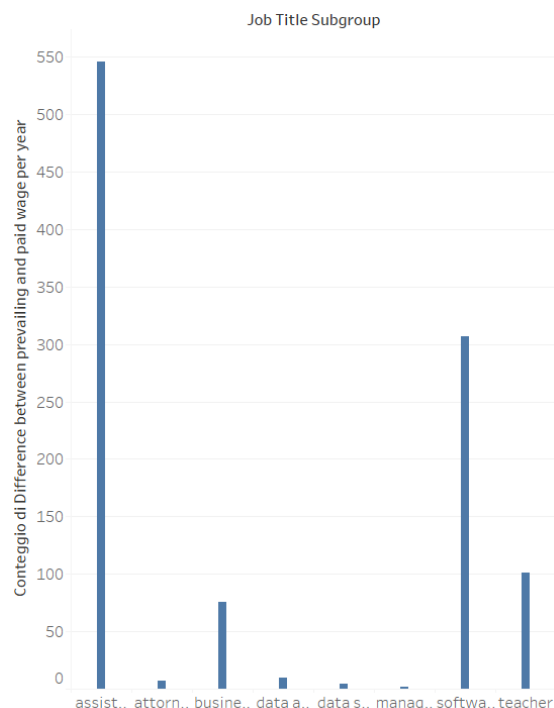
Here we can see, for each job title, the number of employees that are paid less than the prevailing wage per year.

1) The parity price dataset is characterized by a higher level of aggregation for what concerns the states. For this reason, during the blending, we have to select this dataset as "primary data source".
2) For a successful mapping we have to specify that the variables to be connected are "State" and "Work state" (clicking on "Custom").

Now that the 2 datasets are blended, I can use calculation to compute the adjusted paid wage per year. This is the formula I'm going to use:

ADJ paid wage = Median(Paid wage per year) / ATTR(Parity price)

The 5 states I want to consider are: Alabama, California, Main, Massachusetts and Washington. To do that I use the following calculation:

Then I drag-and-drop "States view" into filters and I select "VIEW" only.

- Columns: Job Title Subgroup, State
- Rows: Median(Paid Wage per Year), AGG(ADJ paid wage)

This is the requested plot: