Data Analytics Project 15
Brazil Wages - Clustering

Tristano Galimberti*
School of Informatics, M.Sc. Artificial Intelligence
Università della Svizzera italiana, Switzerland
Email: *galimt@usi.ch

***Abstract*—In this report I outline the problem of clustering the wages of workers in Brazil based on the numerical/categorical features provided in the dataset and the process taken in order to find the clusters from which to make the most meaningful categorizations of the salaries. The report concludes that although it's challenging to obtain exact clusters for the jobs and their respective sectors, we can get a good estimate for which paygrade is most common within each sector and also what type of a bonus a person would expect to get (if at all) at any given salary bracket.**

I. Exploration and description of the data

The dataset provided is a record of the wages received by just over one million workers in São Paulo, Brazil in the month of October, 2017. The attributes in the order that they appear in the dataset, before cleaning, with their respective datatypes and units (if applicable) are as follows:

- **ID** (Discrete), Personal Identification - Each worker is assigned an integer identifier starting from 1.
- **Job** (Categorical, 3461 unique categories), Type of job that the person works.
- **Sector** (Categorical, 90 unique categories), Sector of the above job.
- **Month salary** (Continuous, units: Brazilian Real - R$), The person's regular monthly salary, before tax and other additions/subtractions.
- **13 salary** (Continuous, units: Brazilian Real - R$), in Brazil people occasionally receive an 'extra' 13th payment in December around Christmas time.
- **Eventual Salary** (Continuous, units: Brazilian Real - R$), eventual variable, only non-zero if the person has received extra salary.
- **Indemnity** (Continuous, units: Brazilian Real - R$), eventual variable, security against loss of finances (non-zero values only account for 0.1% of the dataset).
- **Extra Salary** (Continuous, units: Brazilian Real - R$), eventual variable, upon analysis of the data I found that this refers to a month-specific bonus where the employee receives more than they do in their *Total Salary* for the month.
- **Discount Salary** (Continuous, units: Brazilian Real - R$), eventual variable, represents a subtraction from the monthly salary.

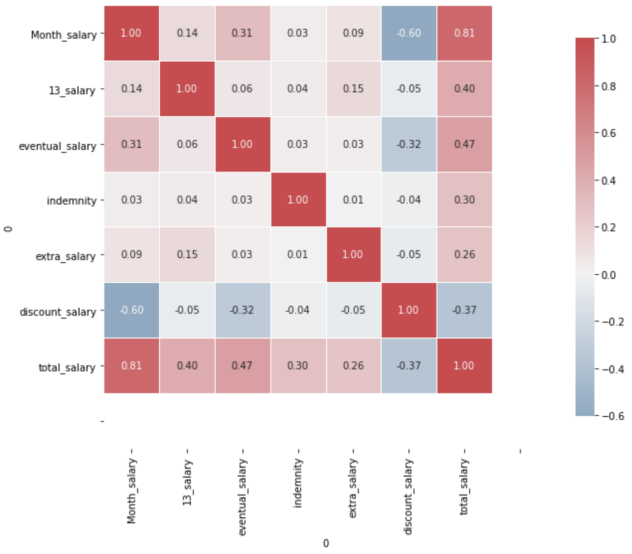

Figure 1. Correlation table for all relevant numerical attributes.

- **Total Salary** (Continuous, units: Brazilian Real - R$), the total salary that a person receives after all the previous fields have and their tax have been taken into account.
- **NaN** (Last field in the csv), continuous

Although this project is entitled *'Brazil Wages'*, we can immediately say that this dataset would not be a good reflection of wages in the entirity of Brazil, given that São Paulo is the largest city in terms of population in Brazil - in fact it has more than double the number of inhabitants as the 2nd largeest city - Rio de Janeiro. For this reason its analysis and predictive power is somewhat limited to the region of São Paulo and perhaps Rio de Janeiro and similarly populated cities, but certainly is not a reliable reflection of wages in the whole of Brazil.

From the correlation table in Figure 1, we can see that there aren't any unexpected positive or negative correlations that we could take advantage of. For example, we would expect that *total salary* and *month salary* have a high correlation because the total salary is simply the monthly salary with tax deducted and occasional bonuses included. Also, it makes sense that *month salary* has a strong negative correlation with *discount salary* because *discount salary* is a salary deduction and is

	Month_salary	13_salary	eventual_salary	indemnity	extra_salary	discount_salary	total_salary
count	1079523.00000	1079523.00000	1079523.00000	1079523.00000	1079523.00000	1079523.00000	1079523.00000
mean	4366.38779	216.78596	135.14495	17.77238	116.75230	-95.58791	3411.43863
std	4410.57808	1094.43926	1252.64814	888.64513	656.44810	1328.77111	3410.96202
min	0.00000	-5298.81000	-18586.41000	-6729.78000	-26393.95000	-109497.23000	-47288.97000
25%	2126.37000	0.00000	0.00000	0.00000	0.00000	0.00000	1668.96000
50%	3389.85000	0.00000	0.00000	0.00000	0.00000	0.00000	2668.14000
75%	5082.74000	0.00000	0.00000	0.00000	96.96000	0.00000	3970.09500
max	131128.28000	86381.12000	155861.78000	327498.03000	115174.83000	0.00000	342151.20000

Figure 2. Standard Statistics

hence given as a negative value, so the bigger the deduction, the smaller the monthly salary will be.

Figure 2 is a table that shows the relevant information taken from the call of `df.describe()` after the pre-processing step outlined in the next section. One thing that can be noted from this table is that the fields **13 salary**, **eventual salary**. **indemnity** and **discount salary** contain a vast majority of zero values, and hence these attributes apply to only a small minority of wages, so we shall concentrate most on the categorical attributes (job and sector) and their relation to **total salary, month salary** and **extra salary** (bonuses).

Displaying the total and monthly salaries for the whole dataset reveals some more details of the distributions that they follow.

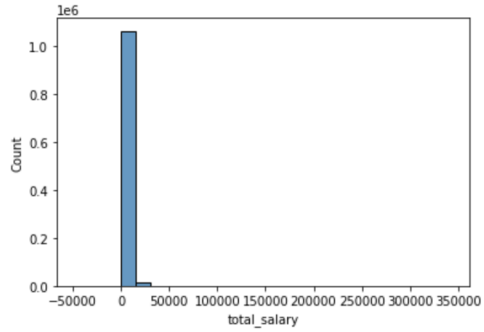

Figure 3. Histogram of total salary

Beyond exposing the wealth gap, Figure 3 is not particularly insightful because it is including the frequency of all salaries up until the highest salary in the city, the consequence this has for the resolution of the chart is that the lower 99% of salaries are below 30,000 Brazilian Real and are hence contained within the first 2 bins. To remedy this we can display the salaries logarithmically to expose the details in the higher salary ranges.

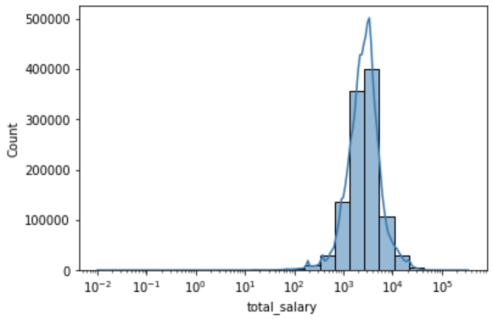

Figure 4. Histogram of total salary with a logarithmically scaled x-axis

Interestingly, plotting the salaries with a logarithmic scale in 4 we can see a Gaussian distribution emerge.

Alternatively, for our purposes it makes more sense to plot monthly salaries as opposed to total salaries because they're more closely related to the types of job. The histogram for this is given in Figure 5.

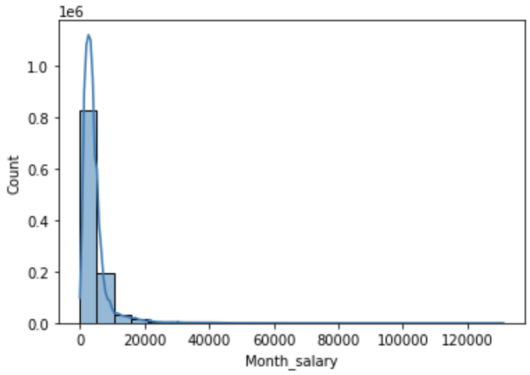

Figure 5. Histogram of monthly salary

Again, the details in the lower salary range suffers because 97% of the workers fall into the first two bins. Figure 6 shows a histogram on the data, where the upper salary bound is capped at 30,000 Brazilian Real. We can remedy this by applying an upper salary limit, for the purposes of viewing the characteristics of the distribution where it matters most - around the mean.

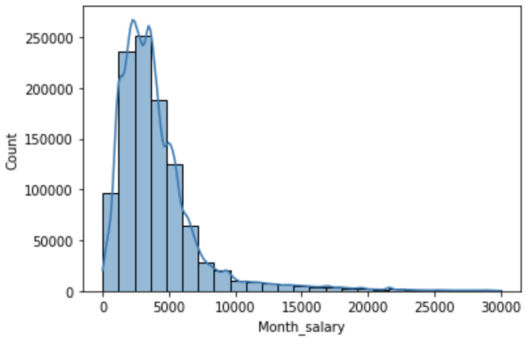

Figure 6. Histogram of monthly salaries up to 30,000R$

From the figure above, we can predict that a clustering algorithm such as K-means will choose for the majority cluster centres to be placed in the 0-30,000R$ range. All of these plots were obtained after the necessary data cleaning.

For this particular analysis, I decided to omit the fields *indemnity* and *discount salary* because they provide minimal insight to the particular salary bracket that somebody could be in and hence, clustering the salaries with respect to these values wouldn't yield a particularly convincing analysis; especially because they're only non-zero in the small minority of cases.

We can also observe the 20 highest paying sectors in Sao Paolo in Figure 7

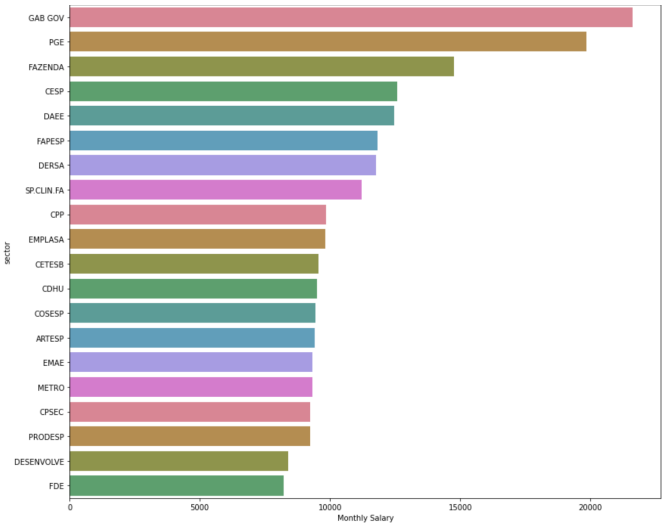

Figure 7. Plot of the 20 highest-paying sectors using the mean of all salaries within that sector

On the other hand, we can also see the lowest paying sectors in Figure 8

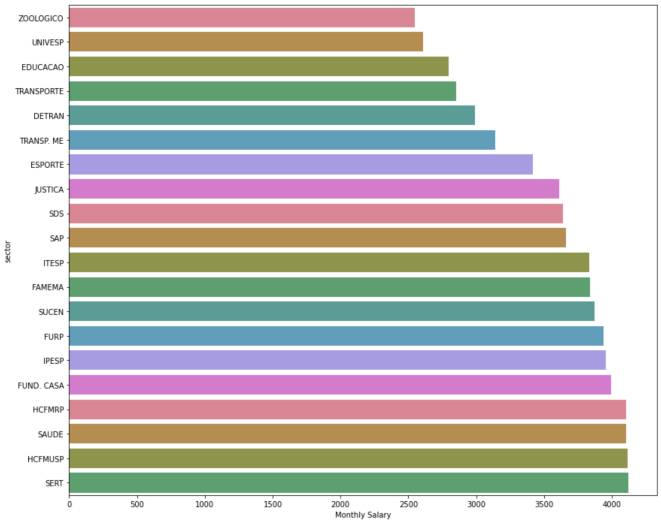

Figure 8. Plot of the 20 lowest-paying sectors using the mean of all salaries within that sector

In future analysis we'll find that it's difficult to properly cluster jobs by salary tier only considering there are thousands of job types in this dataset with an almost complete majority having overlapping pay grades. This allows us to present percentage probabilities that a certain paygrade belongs to a certain job and vice-versa. However we can also perform some analysis by clustering salary tiers by bonus amounts, Figure 9 is a figure which plots the bonus salary against the monthly salary.

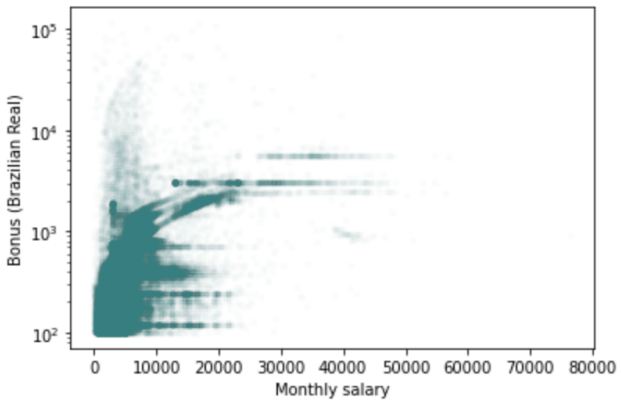

Figure 9. Bonus amount vs. monthly salary with bonus on a log scale

From this plot we notice that there is indeed a weak correlation between bonus amounts and monthly salary tiers. We also notice that there are about seven discrete bonus values that appear much more consistently across all the monthly salary pay grades, shown by the dark horizontal streaks. These are at approximately B$ 120, 250, 350, 800, 2500, 3000 and 6000 and also support the pattern of higher bonus meaning higher salary - for example the B$120 and B$250 bonuses cease to occur at a salary of about B$22,000.

We can also plot the same data on a log-log basis

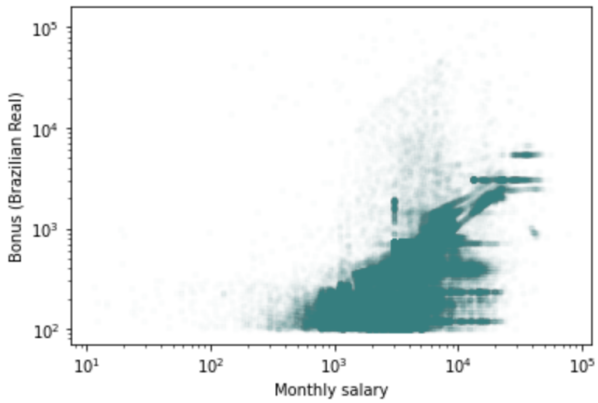

Figure 10. Bonus amount vs. monthly salary with both variables on a log scale

This more clearly outlines the logarithmic nature of the weak correlation between bonus and monthly salary.

II. Data Preparation & Pre-processing

Several steps of pre-processing had to be taken prior to determining that the data was ready to apply the clustering algorithms. A good indicator for this was the initial call of `df.describe()` which outlined minimum values of zero for some fields, for example monthly salary, which were very unlikely/impossible.

There wasn't a description of each field included in the `.csv`, neither was there an attached readme so it took some searching through the data to infer how the field values were related to each other and what they meant exactly, for example: total salary was less than the monthly salary because it took into account tax.

Firstly, the values in the `.csv` file all imported as objects, meaning that both the numerical and categorial fields all had the type `object`, so they first required conversion to their respective datatypes before proceeding with the analysis.

Secondly, in the monthly salary field, there were a small number of entries which were strings, again making numerical analysis plotting of the entire attribute impossible, so I decided to omit the rows corresponding to those values considering it was of little consequence to the 1 million datapoints in the dataset. Further I chose to remove rows that contained elements corresponding to a total salary of zero (5236 datapoints) and a month salary of zero (4333 datapoints);

for **monthly salary** I noticed that the lowest non-zero value was 12%, however there were 22 values which had a humidity of 0. This is physically impossible for atmospheric conditions. Only 22 values out of 96,000 had an invalid (zero) value, so I replaced those values with the mean of the humidity, which was calculated without using those values. The maximum humidity in 2 reads 1.00, which I initially assumed to be another error, however after some research I found there is a meteorogical phenomenon called 'supersaturation' [**?**] which is common in most parts of the world.

As outlined in the notebook, the attribute **NaN** was a column full of zeros, with non-zero values existing only where the **total salary** field contained zeros. Regardless, I decided to exclude this field from the analysis because the name of the attribute was not specified in the `.csv`.

III. Clustering

The prediction model I chose was based on performing K-means on the salary data to find which salary brackets (as defined by the K-means algorithm) most often corresponded to each of the 88 sectors present in the dataset.

When running K-means on the **total salary** data for K in the range from 1 to 25, we get the relationship between WCSS and K value in Figure 11.

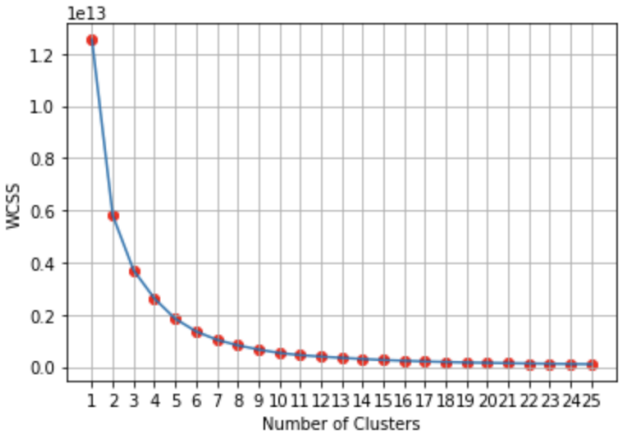

Figure 11. Linear plot of WCSS for K-means for K in range (2-25)

and logarithmically

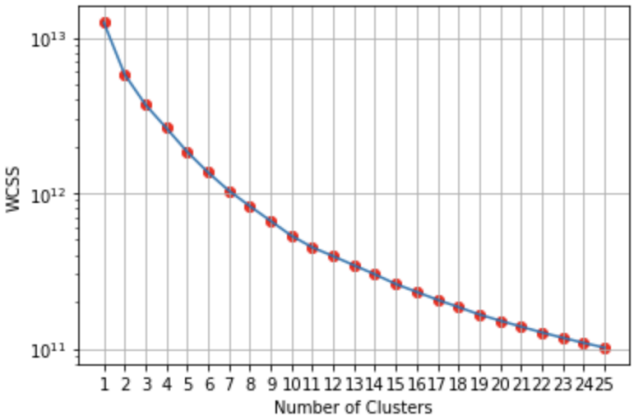

Figure 12. Logarithmic plot of WCSS for K-means for K in range (2-25)

The WCSS is a measure of the error inside each cluster, or (Within-Cluster-Sum-of-Squares) has the formula

$$\text{WCSS} = \sum_{i \in n} (X_i - Y_i)^2 \tag{1}$$

so this inverse relationship holds true when applying K-means to the salary data. We can observe from Figure 11 that we have sharply diminishing returns with respect to the number of clusters. For this reason I chose to use K-means on 9 clusters. The K-means algorithm resulted in the following centroids which are also provided in the notebook.

1	2898.052791
2	20320.142652
3	7713.873899
4	4731.089408
5	65640.530123
6	1337.225002
7	218471.073500
8	12697.108269
9	30879.247853

As expected, the majority of cluster centroids (7/9) were in the 0-30,000R$ range specified earlier. I then applied this K-means algorithm on the **job** (Figure 14) and **sector** (Figure **??**) attributes to view how many workers in each job/sector fell within each cluster.

ADVOGADO 01	1	0	1	7	0	1	0	0	0
ADVOGADO 02	1	0	0	12	0	0	0	0	0
ADVOGADO 03	1	0	1	4	0	0	0	1	0
ADVOGADO 04	1	0	0	5	0	0	0	0	0
ADVOGADO 05	0	0	0	4	0	0	0	1	0
ADVOGADO 06	0	0	0	5	0	0	0	0	0
ADVOGADO 07	3	0	0	2	0	0	0	0	0
ADVOGADO 08	0	0	2	1	0	0	0	0	0
ADVOGADO 09	1	1	2	3	0	0	0	0	0
ADVOGADO 10	1	0	4	3	0	0	0	0	0
ADVOGADO 11	0	0	3	4	0	0	0	1	0
ADVOGADO 12	0	1	2	1	0	0	0	0	0

Figure 13. List of the first 12 jobs and the frequency of their occurence in each cluster

cluster	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
sector									
ADM GERAL	5052	177	2232	3158	2	21282	0	1173	18
AGEM	3	0	4	6	0	5	0	0	0
AGEMCAMP	1	0	3	9	0	1	0	0	0
AGEMVALE	1	0	3	0	0	0	0	0	0
ARSESP	15	1	48	52	0	8	0	12	0
ARTESP	6	0	22	18	0	2	0	4	0
CASA CIVIL	97	2	26	88	0	12	0	9	1
CBPM	17	0	5	5	0	15	0	1	1
CDHU	128	4	169	165	0	23	0	151	0
CEETEPS	6158	36	2158	4329	1	5972	0	553	17

Figure 14. List of the first 12 sectors and the frequency of their occurence in each cluster, where the columns represent the clusters 1-9

The full results for the K-means algorithm for the 3366 unique jobs and 88 unique sectors are provided near the end of the notebook.

Given the format of the dataset, it's difficult to perform 2d clustering due to the fact that we're aiming to cluster categorical data with respect to numerical values. We can however try to find a correlation between salary levels and the existence/size of their respective bonuses and cluster them instead.

Running K-means instead on the **monthly salary** and **extra salary** attributes with $k = 6$, we get the following centroids.

Cluster	Bonus amount	Monthly Salary
1	476.182962	4453.607387
2	1958.167807	15717.981541
3	762.504559	7329.755319
4	8037.034452	19749.995017
5	192.912330	2073.019831
6	89478.049000	11267.65

IV. Evaluation

As mentioned earlier in the report, it was a challenging task to cluster the jobs considering that there are 3388 unique values with a wide range of salaries, but I was able to group each job and each sector into clusters, which could be used to determine the likelihood that any one sector/cluster would exist within a certain pay grade.

After this I also performed K-means to cluster find clusters in the 2 dimensional scatter plot between monthly salaries and bonuses.

These results by no means provide a perfect clustering of all of the jobs, I believe it could be further improved upon in the future with more computational resources by incorporating Gaussian mixture models on the sector-wise monthly salaries.