Task 1

Fabio Burri, Mischa Haenen, Robin Galeazzi, and Timon Galeazzi

2024-04-23

Task 1

Loading Dataset and inital exploration

For the task 1, we started by loading the dataset 'data_wage.RData'. First we download all relevant packages for this task.

Then our idea was to get a quick overview of the dataset. Which dimensions does the dataset have? Which variable types are used?

We have 10'809 observations with 78 variables like gender, age, country, education, ecetera. We were then interested in getting a visual overview of the most important data and the percentages.

Our aim was then to find the dependent variable (Y), in this case 'wage'.

```
names(data)
```

```
##  [1] "gender"
##  [2] "age"
##  [3] "country"
##  [4] "education"
##  [5] "undergraduate_major"
##  [6] "job_role"
##  [7] "industry"
##  [8] "years_experience"
##  [9] "ML_atwork"
## [10] "Activities_Analyze.and.understand.data.to.influence.product.or.business.decisions"
## [11] "Activities_Build.and.or.run.a.machine.learning.service.that.operationally.improves.my.product.
## [12] "Activities_Build.and.or.run.the.data.infrastructure.that.my.business.uses.for.storing..analyzi
## [13] "Activities_Build.prototypes.to.explore.applying.machine.learning.to.new.areas"
## [14] "Activities_Do.research.that.advances.the.state.of.the.art.of.machine.learning"
## [15] "Activities_None.of.these.activities.are.an.important.part.of.my.role.at.work"
## [16] "Notebooks_Kaggle.Kernels"
## [17] "Notebooks_Google.Colab"
## [18] "Notebooks_Azure.Notebook"
## [19] "Notebooks_Google.Cloud.Datalab"
## [20] "Notebooks_JupyterHub.Binder"
## [21] "Notebooks_None"
## [22] "cloud_Google.Cloud.Platform..GCP."
## [23] "cloud_Amazon.Web.Services..AWS."
## [24] "cloud_Microsoft.Azure"
## [25] "cloud_IBM.Cloud"
```

```
## [26] "cloud_Alibaba.Cloud"
## [27] "cloud_I.have.not.used.any.cloud.providers"
## [28] "Programming_Python"
## [29] "Programming_R"
## [30] "Programming_SQL"
## [31] "Programming_Bash"
## [32] "Programming_Java"
## [33] "Programming_Javascript.Typescript"
## [34] "Programming_Visual.Basic.VBA"
## [35] "Programming_C.C.."
## [36] "Programming_MATLAB"
## [37] "Programming_Scala"
## [38] "Programming_Julia"
## [39] "Programming_SAS.STATA"
## [40] "Programming_language_used_most_often"
## [41] "ML_framework_Scikit.Learn"
## [42] "ML_framework_TensorFlow"
## [43] "ML_framework_Keras"
## [44] "ML_framework_PyTorch"
## [45] "ML_framework_Spark.MLlib"
## [46] "ML_framework_H2O"
## [47] "ML_framework_Caret"
## [48] "ML_framework_Xgboost"
## [49] "ML_framework_randomForest"
## [50] "ML_framework_None"
## [51] "Visualization_ggplot2"
## [52] "Visualization_Matplotlib"
## [53] "Visualization_Altair"
## [54] "Visualization_Shiny"
## [55] "Visualization_Plotly"
## [56] "Visualization_None"
## [57] "percent_actively.coding"
## [58] "How.long.have.you.been.writing.code.to.analyze.data."
## [59] "For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school.."
## [60] "Do.you.consider.yourself.to.be.a.data.scientist."
## [61] "data_Categorical.Data"
## [62] "data_Genetic.Data"
## [63] "data_Geospatial.Data"
## [64] "data_Image.Data"
## [65] "data_Numerical.Data"
## [66] "data_Sensor.Data"
## [67] "data_Tabular.Data"
## [68] "data_text.Data"
## [69] "data_Time.Series.Data"
## [70] "data_Video.Data"
## [71] "explainability.model_Examine.individual.model.coefficients"
## [72] "explainability.model_examine.feature.correlations"
## [73] "explainability.model_Examine.feature.importances"
## [74] "explainability.model_Create.partial.dependence.plots"
## [75] "explainability.model_LIME.functions"
## [76] "explainability.model_SHAP.functions"
## [77] "explainability.model_None.I.do.not.use.these.model.explanation.techniques"
## [78] "wage"
```

```
summary(data$wage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    6811   34780   53048   75687  551774
```

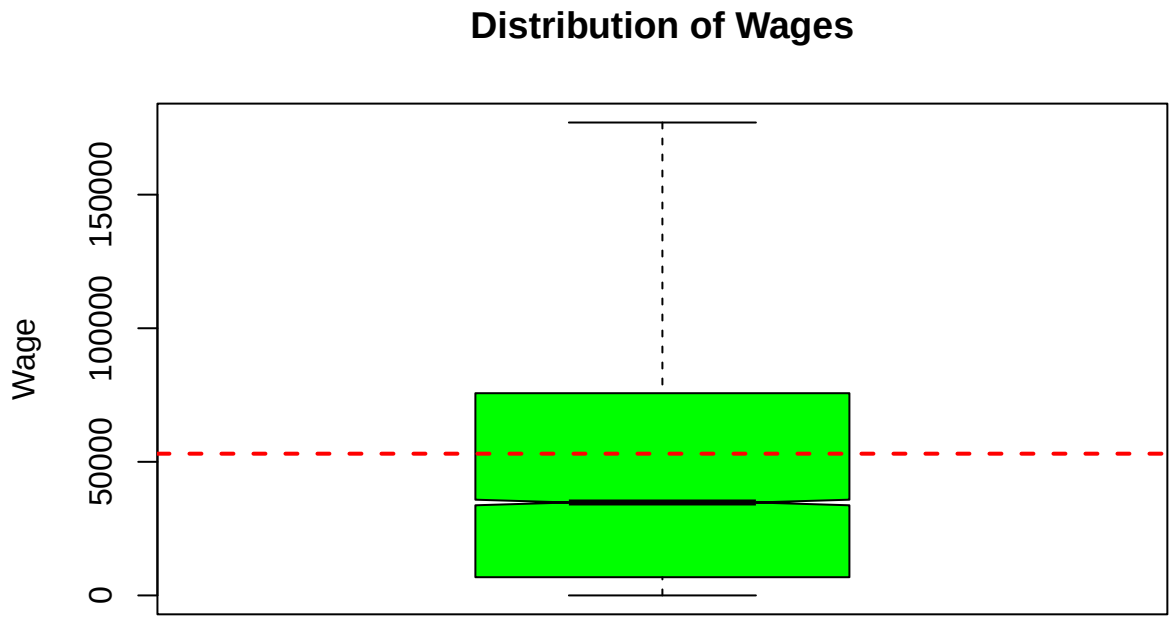
We can see that the average salary is over 53,000 dollars. As the average wage is heavily distorted by high salaries - the maximum is 552,000 dollars - the median of just under 35,000 dollars is more meaningful. Because the minimum value is zero, we were interested to see how many people stated 0 as their salary (e.g. students).

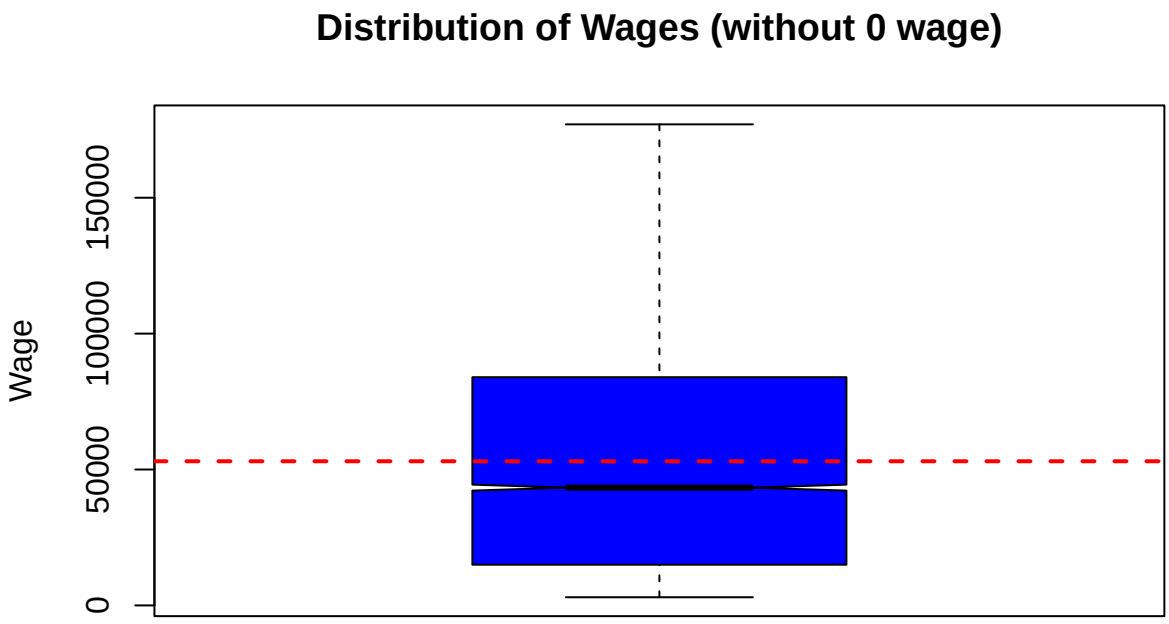
```
## [1] 1000
```

This amounted to a total of 1000 people.We then created another dataset without the people who entered 0 in the wage.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3007   14983   43342   58456   83968  551774
```

The average value is 58500, the media just under 43500.Additionally, we would like to show the distribution of wages graphically.

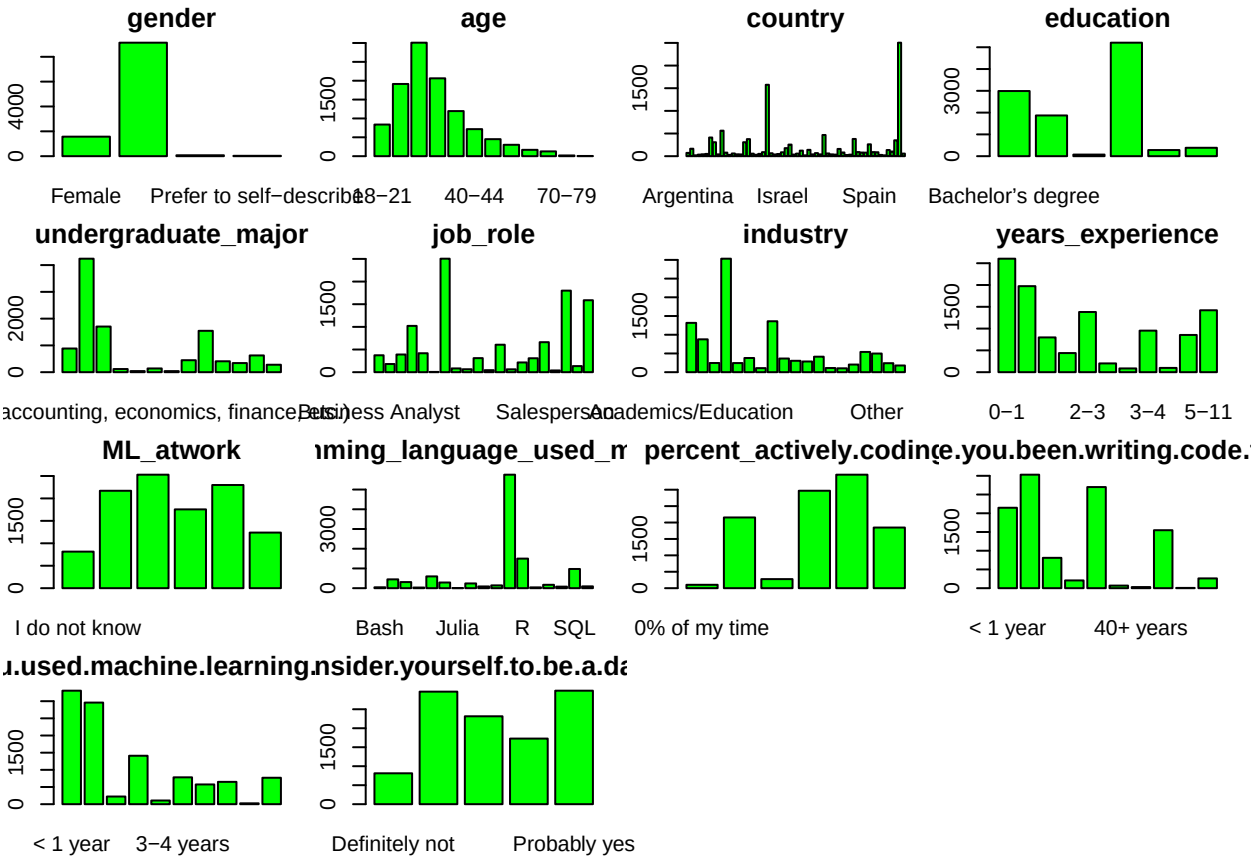




We will then analyze the categorical and numerical values separately. To do this, we save the respective data as new values. Our aim was then to find out the number of numerical and categorical values.

```
## [1] 64
```

```
## [1] 14
```


gender
0
4000
Female
Prefer to self−describe
age
0
1500
18−21
40−44
70−79
country
0
1500
Argentina
Israel
Spain
education
0
3000
Bachelor's degree
undergraduate_major
0
2000
accounting, economics, finance (etc.)
job_role
0
1500
Business Analyst
Salesperson
industry
0
1500
Academics/Education
Other
years_experience
0
1500
0−1
2−3
3−4
5−11
ML_atwork
0
1500
I do not know
ıming_language_used_m
0
3000
Bash
Julia
R
SQL
percent_actively.codin
0
1500
0% of my time
e.you.been.writing.code.
0
1500
< 1 year
40+ years
ı.used.machine.learning.
0
1500
< 1 year
3−4 years
nsider.yourself.to.be.a.da
0
1500
Definitely not
Probably yes