Salary Prediction by Machine Learning

Zhihan Xian

2021-10-28

# 1 Summary/Abstract

*Write a summary of your project.*

# 2 Introduction

The dataset I obtained is about jobs and salaries. The dataset contains variables including: - age group - industry - job title - salary - country - state - years of professional experience - total years of experience - education level - gender - race

Most variables are categorical variables. Some of them are ordinal, such as age group, years of experience, which are categorical in nature but also have comparative order. The only numeric data is the outcome of interest, which is salary.

dataset source: <https://oscarbaruffa.com/messy/>

## 2.1 Questions/Hypotheses to be addressed

* Is it possible to predict a job’s salary based on the variables mentioned above?
* What are the major determinants of a job’s salary? Is there a good model summarizing the dataset?

# 3 Methods and Results

*In most research papers, results and methods are separate. You can combine them here if you find it easier. You are also welcome to structure things such that those are separate sections.*

## 3.1 Data import and cleaning

See processingscript.R file in code/processingz-code directory. Briefly, after loaded the data from the xls file, data related to United States were extracted (Country = US and Currency = USD). Then, since very few rows contains NA entry, these rows are removed from the dataset. Then, for each variable, I did some plotting to see its structure. Most of them only contains a few entry types and are ready to go, while some variables contain messy text entries that have to be cleaned up. For the industry column, I picked the top 20 industries since only top 20 industries have enough datapoints for model training and testing. For the job title text entry, it is too messy to clean up since people may call a same position with different names. Therefore, this variable is dropped. I filtered out observations where salary values are less than $10,000, since it is very unlikely. For the variable race, around 95% of entry are white. Since this variable has little variation, it is dropped from the analysis too.

## 3.2 Exploratory analysis

*Use a combination of text/tables/figures to explore and describe your data. You should produce plots or tables or other summary quantities for the most interesting/important quantities in your data. Depending on the total number of variables in your dataset, explore all or some of the others. FIgures produced here might be histograms or density plots, correlation plots, etc. Tables might summarize your data.*

*Continue by creating plots or tables of the outcome(s) of interest and the predictor/exposure/input variables you are most interested in. If your dataset is small, you can do that for all variables. Plots produced here can be scatterplots, boxplots, violinplots, etc. Tables can be simple 2x2 tables or larger ones.*

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. t-tests, simple regression model with 1 predictor, etc.) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

Table 3.1 shows a table summarizing the data.

Table 3.1: Data summary table.

|  |  |  |
| --- | --- | --- |
|  | Height | Weight |
| Min. | 133.00 | 45.00 |
| 1st Qu. | 155.25 | 54.25 |
| Median | 166.00 | 73.00 |
| Mean | 165.50 | 72.00 |
| 3rd Qu. | 177.25 | 87.50 |
| Max. | 192.00 | 110.00 |

Figure 3.1 shows the messy data structure for the variable Industry.

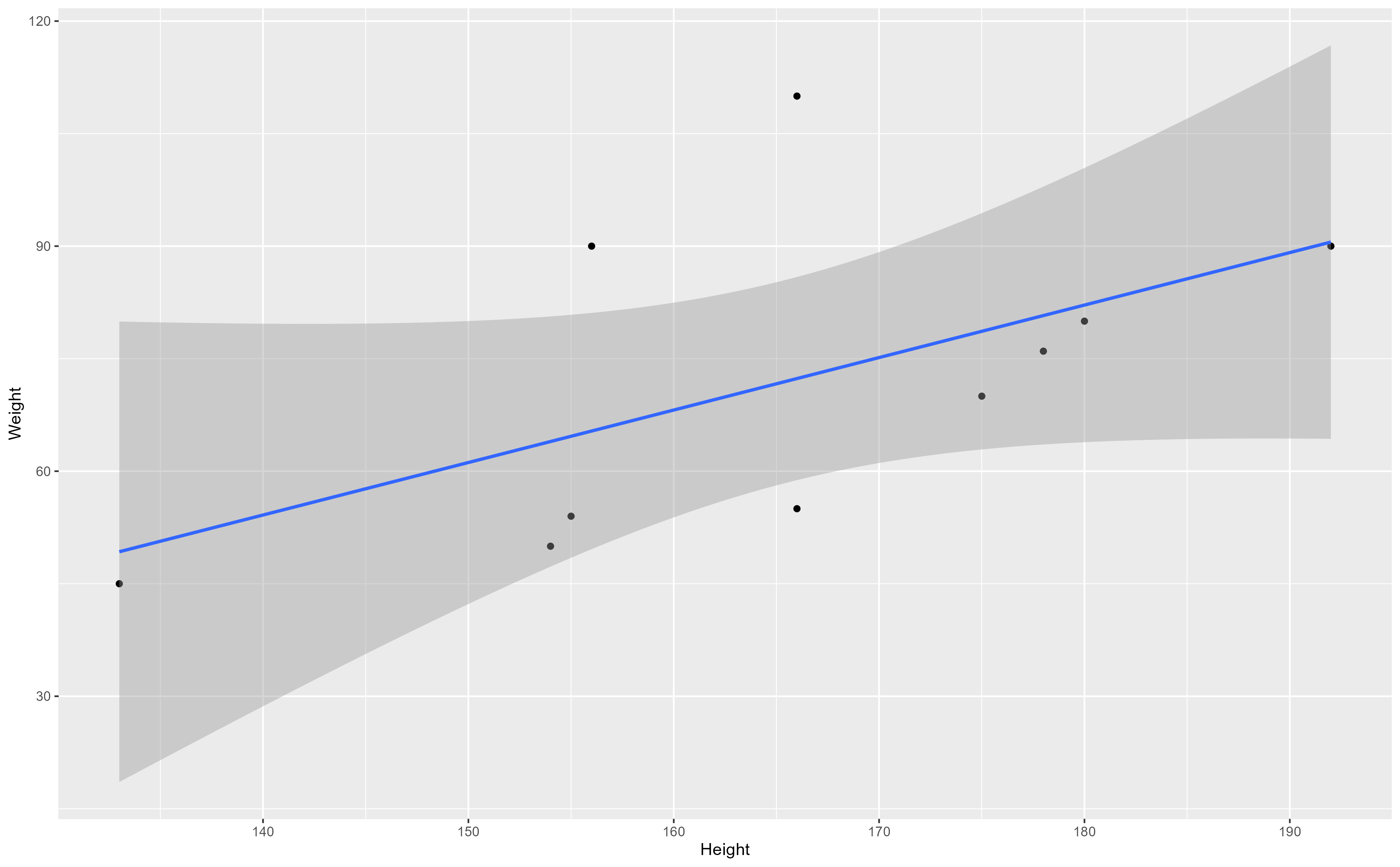


Figure 3.1: Analysis figure.

## 3.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example table 3.2 shows a table summarizing a linear model fit.

Table 3.2: Linear model fit table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -43.7883068 | 61.1150617 | -0.7164896 | 0.4940713 |
| Height | 0.6996272 | 0.3675692 | 1.9033889 | 0.0934786 |

# 4 Discussion

## 4.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 4.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 4.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# 5 References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>