Wage Prediction Using Machine Learning and AutoML

Feuz Dana Livia, Hoogstrate Ian, Kuchen Rahel, Vandecruys Piet

1. Load Required Libraries

We start by loading in the necessary libraries. These libraries are for data exploration machine learning autoML and some require extra installations. For autoML you need to have the 64 bit version of JAVA.

```
# Install missing packages
packages <- c("caret", "randomForest", "pdp", "ggplot2", "dplyr", "caretEnsemble", "h2o", "data.table")
to_install <- packages[!packages %in% installed.packages()[, "Package"]]
if(length(to_install)) install.packages(to_install)


# Load libraries
library(h2o)
library(caret)
library(caretEnsemble)
library(randomForest)
library(pdp)
library(ggplot2)
library(dplyr)
library(data.table)
```

2. Load the Data

We load a preprocessed dataset data_wage.RData which contains the wages and all the features that we will need for the prediction.

```
load("data_wage.RData")
df <- data
```

3. Data Preprocessing

Goal: Understand the data before modeling. Why? Data-driven decisions start with exploration.

We will first take a look at our data, we can see that our data exists out of more than 10 thousand rows and 78 features. These features include, age, years of experience, industry job role, if they have used ML before and a whole lot more. We then check for missing values, and we can see that there are none, if there were any missing values, we would remove the row.

```
str(df)
```

```
## 'data.frame':    10809 obs. of  78 variables:
##  $ gender
##  $ age
##  $ country
##  $ education
##  $ undergraduate_major
##  $ job_role
##  $ industry
##  $ years_experience
##  $ ML_atwork
##  $ Activities_Analyze.and.understand.data.to.influence.product.or.business.decisions
##  $ Activities_Build.and.or.run.a.machine.learning.service.that.operationally.improves.my.product.or.w
##  $ Activities_Build.and.or.run.the.data.infrastructure.that.my.business.uses.for.storing..analyzing.
##  $ Activities_Build.prototypes.to.explore.applying.machine.learning.to.new.areas
##  $ Activities_Do.research.that.advances.the.state.of.the.art.of.machine.learning
##  $ Activities_None.of.these.activities.are.an.important.part.of.my.role.at.work
##  $ Notebooks_Kaggle.Kernels
##  $ Notebooks_Google.Colab
##  $ Notebooks_Azure.Notebook
##  $ Notebooks_Google.Cloud.Datalab
##  $ Notebooks_JupyterHub.Binder
##  $ Notebooks_None
##  $ cloud_Google.Cloud.Platform..GCP.
##  $ cloud_Amazon.Web.Services..AWS.
##  $ cloud_Microsoft.Azure
##  $ cloud_IBM.Cloud
##  $ cloud_Alibaba.Cloud
##  $ cloud_I.have.not.used.any.cloud.providers
##  $ Programming_Python
##  $ Programming_R
##  $ Programming_SQL
##  $ Programming_Bash
##  $ Programming_Java
##  $ Programming_Javascript.Typescript
##  $ Programming_Visual.Basic.VBA
##  $ Programming_C.C..
##  $ Programming_MATLAB
##  $ Programming_Scala
##  $ Programming_Julia
##  $ Programming_SAS.STATA
##  $ Programming_language_used_most_often
##  $ ML_framework_Scikit.Learn
##  $ ML_framework_TensorFlow
##  $ ML_framework_Keras
##  $ ML_framework_PyTorch
##  $ ML_framework_Spark.MLlib
##  $ ML_framework_H2O
##  $ ML_framework_Caret
##  $ ML_framework_Xgboost
##  $ ML_framework_randomForest
##  $ ML_framework_None
##  $ Visualization_ggplot2
##  $ Visualization_Matplotlib
##  $ Visualization_Altair
```

```
##  $ Visualization_Shiny
##  $ Visualization_Plotly
##  $ Visualization_None
##  $ percent_actively.coding
##  $ How.long.have.you.been.writing.code.to.analyze.data.
##  $ For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school..
##  $ Do.you.consider.yourself.to.be.a.data.scientist.
##  $ data_Categorical.Data
##  $ data_Genetic.Data
##  $ data_Geospatial.Data
##  $ data_Image.Data
##  $ data_Numerical.Data
##  $ data_Sensor.Data
##  $ data_Tabular.Data
##  $ data_text.Data
##  $ data_Time.Series.Data
##  $ data_Video.Data
##  $ explainability.model_Examine.individual.model.coefficients
##  $ explainability.model_examine.feature.correlations
##  $ explainability.model_Examine.feature.importances
##  $ explainability.model_Create.partial.dependence.plots
##  $ explainability.model_LIME.functions
##  $ explainability.model_SHAP.functions
##  $ explainability.model_None.I.do.not.use.these.model.explanation.techniques
##  $ wage
```

```r
summary(df)
```

```
##                      gender          age                             country
##  Female                 :1571   25-29  :3008   United States of America:2505
##  Male                   :9135   30-34  :2064   India                   :1576
##  Prefer not to say      :  72   22-24  :1914   China                   : 563
##  Prefer to self-describe:  31   35-39  :1195   Other                   : 468
##                                 18-21  : 838   Brazil                  : 412
##                                 40-44  : 717   Russia                  : 380
##                                 (Other):1073   (Other)                 :4905
##                                                                  education
##  Bachelor’s degree                                                   :2990
##  Doctoral degree                                                     :1869
##  I prefer not to answer                                              :  74
##  Master’s degree                                                     :5209
##  Professional degree                                                 : 281
##  Some college/university study without earning a bachelor’s degree: 386
##
##                                                        undergraduate_major
##  Computer science (software engineering, etc.)                  :4239
##  Engineering (non-computer focused)                             :1704
##  Mathematics or statistics                                      :1545
##  A business discipline (accounting, economics, finance, etc.): 884
##  Physics or astronomy                                           : 626
##  Information technology, networking, or system administration: 447
##  (Other)                                                        :1364
##               job_role                                          industry
##  Data Scientist    :2505   Computers/Technology                     :3032
```

```
## Software Engineer :1800   I am a student                          :1361
## Student            :1588   Academics/Education                     :1317
## Data Analyst       :1022   Accounting/Finance                      : 878
## Research Scientist: 662   Online Service/Internet-based Services: 541
## Other              : 606   Other                                   : 498
## (Other)            :2626   (Other)                                 :3182
## years_experience
## 0-1    :2604
## 1-2    :1974
## 5-11   :1421
## 2-3    :1381
## 3-4    : 953
## 4-5    : 854
## (Other):1622
##                                                                                          ML_atwork
## I do not know                                                                             : 815
## No (we do not use ML methods)                                                             :2171
## We are exploring ML methods (and may one day put a model into production)                :2529
## We have well established ML methods (i.e., models in production for more than 2 years)   :1756
## We recently started using ML methods (i.e., models in production for less than 2 years)  :2299
## We use ML methods for generating insights (but do not put working models into production):1239
##
## Activities_Analyze.and.understand.data.to.influence.product.or.business.decisions
## Min.   :0.000
## 1st Qu.:0.000
## Median :1.000
## Mean   :0.541
## 3rd Qu.:1.000
## Max.   :1.000
##
## Activities_Build.and.or.run.a.machine.learning.service.that.operationally.improves.my.product.or.wor
## Min.   :0.0000
## 1st Qu.:0.0000
## Median :0.0000
## Mean   :0.3128
## 3rd Qu.:1.0000
## Max.   :1.0000
##
## Activities_Build.and.or.run.the.data.infrastructure.that.my.business.uses.for.storing..analyzing..an
## Min.   :0.0000
## 1st Qu.:0.0000
## Median :0.0000
## Mean   :0.3126
## 3rd Qu.:1.0000
## Max.   :1.0000
##
## Activities_Build.prototypes.to.explore.applying.machine.learning.to.new.areas
## Min.   :0.0000
## 1st Qu.:0.0000
## Median :0.0000
## Mean   :0.4266
## 3rd Qu.:1.0000
## Max.   :1.0000
##
```

```
##  Activities_Do.research.that.advances.the.state.of.the.art.of.machine.learning
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.259
##  3rd Qu.:1.000
##  Max.   :1.000
##
##  Activities_None.of.these.activities.are.an.important.part.of.my.role.at.work
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.154
##  3rd Qu.:0.000
##  Max.   :1.000
##
##  Notebooks_Kaggle.Kernels Notebooks_Google.Colab Notebooks_Azure.Notebook
##  Min.   :0.0000           Min.   :0.0000         Min.   :0.0000
##  1st Qu.:0.0000           1st Qu.:0.0000         1st Qu.:0.0000
##  Median :0.0000           Median :0.0000         Median :0.0000
##  Mean   :0.3335           Mean   :0.1944         Mean   :0.0741
##  3rd Qu.:1.0000           3rd Qu.:0.0000         3rd Qu.:0.0000
##  Max.   :1.0000           Max.   :1.0000         Max.   :1.0000
##
##  Notebooks_Google.Cloud.Datalab Notebooks_JupyterHub.Binder Notebooks_None
##  Min.   :0.00000                Min.   :0.0000              Min.   :0.0000
##  1st Qu.:0.00000                1st Qu.:0.0000              1st Qu.:0.0000
##  Median :0.00000                Median :0.0000              Median :0.0000
##  Mean   :0.07327                Mean   :0.2774              Mean   :0.3796
##  3rd Qu.:0.00000                3rd Qu.:1.0000              3rd Qu.:1.0000
##  Max.   :1.00000                Max.   :1.0000              Max.   :1.0000
##
##  cloud_Google.Cloud.Platform..GCP. cloud_Amazon.Web.Services..AWS.
##  Min.   :0.0000                    Min.   :0.0000
##  1st Qu.:0.0000                    1st Qu.:0.0000
##  Median :0.0000                    Median :0.0000
##  Mean   :0.2756                    Mean   :0.4596
##  3rd Qu.:1.0000                    3rd Qu.:1.0000
##  Max.   :1.0000                    Max.   :1.0000
##
##  cloud_Microsoft.Azure cloud_IBM.Cloud   cloud_Alibaba.Cloud
##  Min.   :0.0000        Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000        1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000        Median :0.00000   Median :0.00000
##  Mean   :0.2329        Mean   :0.06874   Mean   :0.02692
##  3rd Qu.:0.0000        3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000        Max.   :1.00000   Max.   :1.00000
##
##  cloud_I.have.not.used.any.cloud.providers Programming_Python Programming_R
##  Min.   :0.0000                            Min.   :0.0000     Min.   :0.0000
##  1st Qu.:0.0000                            1st Qu.:1.0000     1st Qu.:0.0000
##  Median :0.0000                            Median :1.0000     Median :0.0000
##  Mean   :0.3209                            Mean   :0.8832     Mean   :0.4208
##  3rd Qu.:1.0000                            3rd Qu.:1.0000     3rd Qu.:1.0000
```

```
##  Max.   :1.0000                            Max.   :1.0000     Max.   :1.0000
##
##  Programming_SQL  Programming_Bash Programming_Java
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000   Median :0.0000
##  Mean   :0.5478   Mean   :0.1929   Mean   :0.2363
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##  Programming_Javascript.Typescript Programming_Visual.Basic.VBA
##  Min.   :0.000                     Min.   :0.0000
##  1st Qu.:0.000                     1st Qu.:0.0000
##  Median :0.000                     Median :0.0000
##  Mean   :0.212                     Mean   :0.0841
##  3rd Qu.:0.000                     3rd Qu.:0.0000
##  Max.   :1.000                     Max.   :1.0000
##
##  Programming_C.C.. Programming_MATLAB Programming_Scala Programming_Julia
##  Min.   :0.0000    Min.   :0.0000     Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000    1st Qu.:0.0000     1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000    Median :0.0000     Median :0.00000   Median :0.00000
##  Mean   :0.2496    Mean   :0.1548     Mean   :0.05791   Mean   :0.01508
##  3rd Qu.:0.0000    3rd Qu.:0.0000     3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000    Max.   :1.0000     Max.   :1.00000   Max.   :1.00000
##
##  Programming_SAS.STATA Programming_language_used_most_often
##  Min.   :0.00000       Python :5754
##  1st Qu.:0.00000       R      :1500
##  Median :0.00000       SQL    : 973
##  Mean   :0.06994       Java   : 598
##  3rd Qu.:0.00000       C/C++  : 447
##  Max.   :1.00000       C#/.NET: 309
##                        (Other):1228
##  ML_framework_Scikit.Learn ML_framework_TensorFlow ML_framework_Keras
##  Min.   :0.000             Min.   :0.0000          Min.   :0.0000
##  1st Qu.:0.000             1st Qu.:0.0000          1st Qu.:0.0000
##  Median :1.000             Median :1.0000          Median :0.0000
##  Mean   :0.702             Mean   :0.5701          Mean   :0.4681
##  3rd Qu.:1.000             3rd Qu.:1.0000          3rd Qu.:1.0000
##  Max.   :1.000             Max.   :1.0000          Max.   :1.0000
##
##  ML_framework_PyTorch ML_framework_Spark.MLlib ML_framework_H2O
##  Min.   :0.0000       Min.   :0.0000           Min.   :0.00000
##  1st Qu.:0.0000       1st Qu.:0.0000           1st Qu.:0.00000
##  Median :0.0000       Median :0.0000           Median :0.00000
##  Mean   :0.2163       Mean   :0.1384           Mean   :0.08844
##  3rd Qu.:0.0000       3rd Qu.:0.0000           3rd Qu.:0.00000
##  Max.   :1.0000       Max.   :1.0000           Max.   :1.00000
##
##  ML_framework_Caret ML_framework_Xgboost ML_framework_randomForest
##  Min.   :0.0000     Min.   :0.0000       Min.   :0.0000
##  1st Qu.:0.0000     1st Qu.:0.0000       1st Qu.:0.0000
##  Median :0.0000     Median :0.0000       Median :0.0000
```

```
##  Mean   :0.1453     Mean   :0.3318       Mean   :0.3482
##  3rd Qu.:0.0000     3rd Qu.:1.0000       3rd Qu.:1.0000
##  Max.   :1.0000     Max.   :1.0000       Max.   :1.0000
##
##  ML_framework_None Visualization_ggplot2 Visualization_Matplotlib
##  Min.   :0.000     Min.   :0.0000        Min.   :0.0000
##  1st Qu.:0.000     1st Qu.:0.0000        1st Qu.:0.0000
##  Median :0.000     Median :0.0000        Median :1.0000
##  Mean   :0.118     Mean   :0.4739        Mean   :0.7495
##  3rd Qu.:0.000     3rd Qu.:1.0000        3rd Qu.:1.0000
##  Max.   :1.000     Max.   :1.0000        Max.   :1.0000
##
##  Visualization_Altair Visualization_Shiny Visualization_Plotly
##  Min.   :0.00000      Min.   :0.0000      Min.   :0.0000
##  1st Qu.:0.00000      1st Qu.:0.0000      1st Qu.:0.0000
##  Median :0.00000      Median :0.0000      Median :0.0000
##  Mean   :0.01397      Mean   :0.1777      Mean   :0.3432
##  3rd Qu.:0.00000      3rd Qu.:0.0000      3rd Qu.:1.0000
##  Max.   :1.00000      Max.   :1.0000      Max.   :1.0000
##
##  Visualization_None          percent_actively.coding
##  Min.   :0.00000    0% of my time        : 103
##  1st Qu.:0.00000    1% to 25% of my time :2155
##  Median :0.00000    100% of my time      : 276
##  Mean   :0.07429    25% to 49% of my time:2969
##  3rd Qu.:0.00000    50% to 74% of my time:3458
##  Max.   :1.00000    75% to 99% of my time:1848
##
##                     How.long.have.you.been.writing.code.to.analyze.data.
##  1-2 years                                  :3030
##  3-5 years                                  :2700
##  < 1 year                                   :2147
##  5-10 years                                 :1548
##  10-20 years                                : 810
##  I have never written code but I want to learn: 261
##  (Other)                                    : 313
##                              For.how.many.years.have.you.used.machine.learning.methods..at.work.or.i
##  < 1 year                                                          :3306
##  1-2 years                                                         :2960
##  2-3 years                                                         :1411
##  3-4 years                                                         : 782
##  I have never studied machine learning but plan to learn in the future: 770
##  5-10 years                                                        : 650
##  (Other)                                                           : 930
##  Do.you.consider.yourself.to.be.a.data.scientist. data_Categorical.Data
##  Definitely not: 813                              Min.   :0.0000
##  Definitely yes:2964                              1st Qu.:0.0000
##  Maybe         :2315                              Median :0.0000
##  Probably not  :1728                              Mean   :0.4823
##  Probably yes  :2989                              3rd Qu.:1.0000
##                                                   Max.   :1.0000
##
##  data_Genetic.Data data_Geospatial.Data data_Image.Data  data_Numerical.Data
##  Min.   :0.00000   Min.   :0.0000       Min.   :0.0000   Min.   :0.0000
```

```
##  1st Qu.:0.00000   1st Qu.:0.0000       1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000       Median :0.0000   Median :1.0000
##  Mean   :0.06088   Mean   :0.1416       Mean   :0.2884   Mean   :0.6337
##  3rd Qu.:0.00000   3rd Qu.:0.0000       3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.0000       Max.   :1.0000   Max.   :1.0000
##
##  data_Sensor.Data data_Tabular.Data data_text.Data   data_Time.Series.Data
##  Min.   :0.0000   Min.   :0.0000    Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000    1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000    Median :1.0000   Median :0.0000
##  Mean   :0.1528   Mean   :0.4421    Mean   :0.5022   Mean   :0.4699
##  3rd Qu.:0.0000   3rd Qu.:1.0000    3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000    Max.   :1.0000   Max.   :1.0000
##
##  data_Video.Data  explainability.model_Examine.individual.model.coefficients
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.0779   Mean   :0.2268
##  3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
##
##  explainability.model_examine.feature.correlations
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3369
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
##  explainability.model_Examine.feature.importances
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.372
##  3rd Qu.:1.000
##  Max.   :1.000
##
##  explainability.model_Create.partial.dependence.plots
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1192
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
##  explainability.model_LIME.functions explainability.model_SHAP.functions
##  Min.   :0.00000                     Min.   :0.00000
##  1st Qu.:0.00000                     1st Qu.:0.00000
##  Median :0.00000                     Median :0.00000
##  Mean   :0.05995                     Mean   :0.04663
##  3rd Qu.:0.00000                     3rd Qu.:0.00000
##  Max.   :1.00000                     Max.   :1.00000
##
```

```
##  explainability.model_None.I.do.not.use.these.model.explanation.techniques
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.101
##  3rd Qu.:0.000
##  Max.   :1.000
##
##       wage
##  Min.   :     0
##  1st Qu.:  6811
##  Median : 34780
##  Mean   : 53048
##  3rd Qu.: 75687
##  Max.   :551774
##
```

```
# Missing data summary
missing_summary <- sapply(df, function(x) sum(is.na(x)))
print(missing_summary)
```

```
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##                                              Activities_Analyze.and.understand.data.to.influence.produc
##
##                    Activities_Build.and.or.run.a.machine.learning.service.that.operationally.improves.
##
## Activities_Build.and.or.run.the.data.infrastructure.that.my.business.uses.for.storing..analyzing..an
##
##                                                  Activities_Build.prototypes.to.explore.applying.machin
##
##                                                  Activities_Do.research.that.advances.the.state.of.the.
##
##                                                   Activities_None.of.these.activities.are.an.important.
##
##                                                                                                     N
##
##
```

```
##
##                                                                                                   No
##
##                                                                                              Noteboo
##
##                                                                                                 Note
##
##
##
##                                                                                          cloud_Googl
##
##                                                                                            cloud_Ama
##
##
##
##
##
##
##
##                                                                                    cloud_I.have.not.us
##
##
##
##
##
##
##
##
##
##
##
##                                                                                           Programmin
##
##                                                                                               Progra
##
##
##
##
##
##
##
##
##
##
##
##                                                                                       Programming_la
##
##                                                                                                  ML_
##
##                                                                                                    I
##
##
##
##
```

```
##
##                                                                                                    MI
##
##
##
##
##
##
##
##                                                                                                   ML
##
##
##
##
##
##                                                                                                    Vi
##
##
##
##
##
##
##
##
##
##                                                                                                     I
##
##                                                                           How.long.have.you.been.writing
##
##                                         For.how.many.years.have.you.used.machine.learning.methods.
##
##                                                                               Do.you.consider.yourself.
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##                                                                      explainability.model_Examine.indivi
```

```
##
##                                                           explainability.model_exami
##
##                                                            explainability.model_Exam
##
##                                                        explainability.model_Create.pa
##
##                                                                        explainabili
##
##                                                                        explainabili
##
##                                     explainability.model_None.I.do.not.use.these.model
##
##
##
```

```r
# Decision: Drop rows with missing data if the proportion is low.
cat("Rows before NA removal:", nrow(df), "\n")
```

```
## Rows before NA removal: 10809
```

```r
df <- na.omit(df)
cat("Rows after NA removal:", nrow(df), "\n")
```

```
## Rows after NA removal: 10809
```

4 Data exploration

Why? To justify feature selection based on data relationships.

While exploring the data, we noticed some unusual cases—specifically, individuals aged 18 to 21 reporting extremely high incomes. This raised the question: are these legitimate data points, or are they outliers?

To investigate, we filtered the data to include only those earning over $100,000. Typically, people this young or with limited experience rarely reach that income level. For instance, it's unlikely that someone under 25 or with less than 3 years of experience would be earning over $100,000.

To help identify potential anomalies, we flagged all individuals meeting these criteria (earning over $100K and either under 25 years old or with less than 3 years of experience) as possible outliers. This made it easier to track and analyze these cases separately.

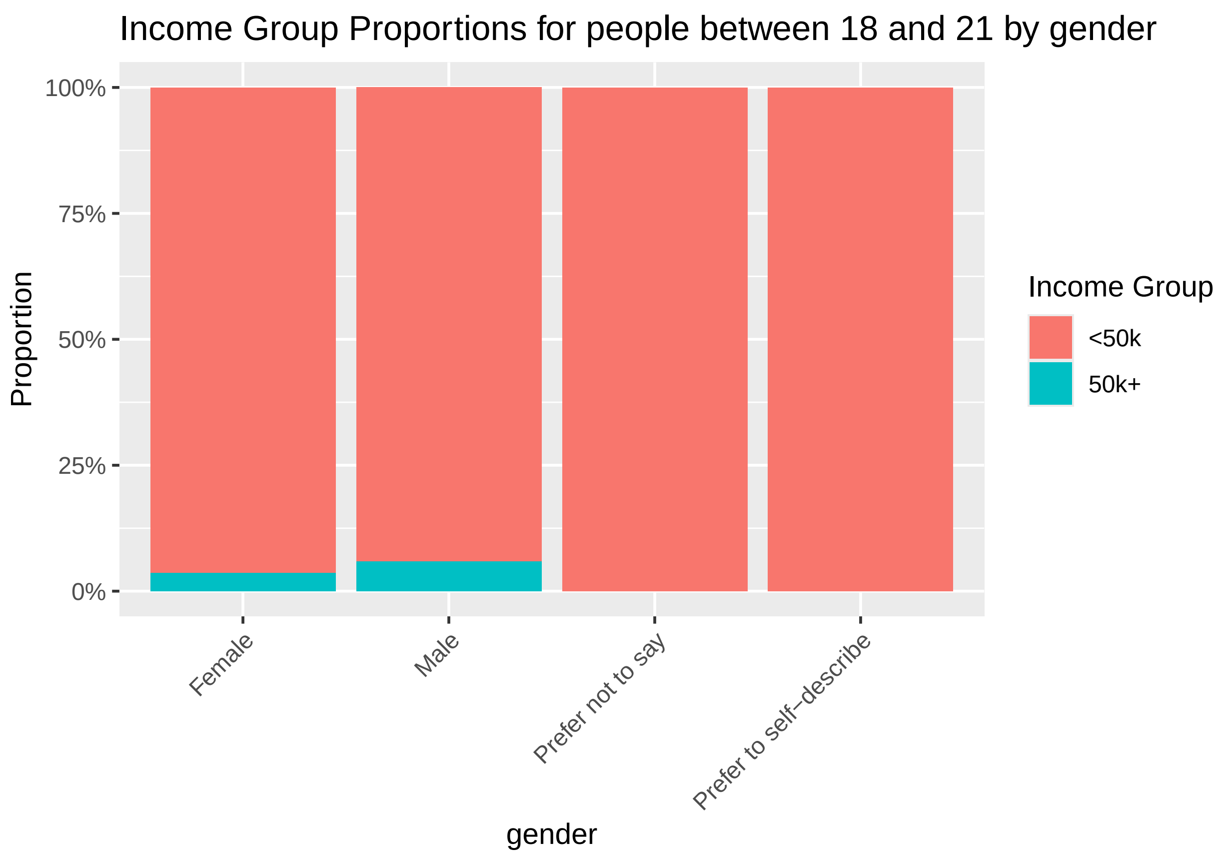
```r
# Outlier flags
df$outlier_flag <- with(df,
                        (wage > 100000 & (years_experience %in% c("0-1", "1-2") | age %in% c("18-21", "2
)
View(df)
```

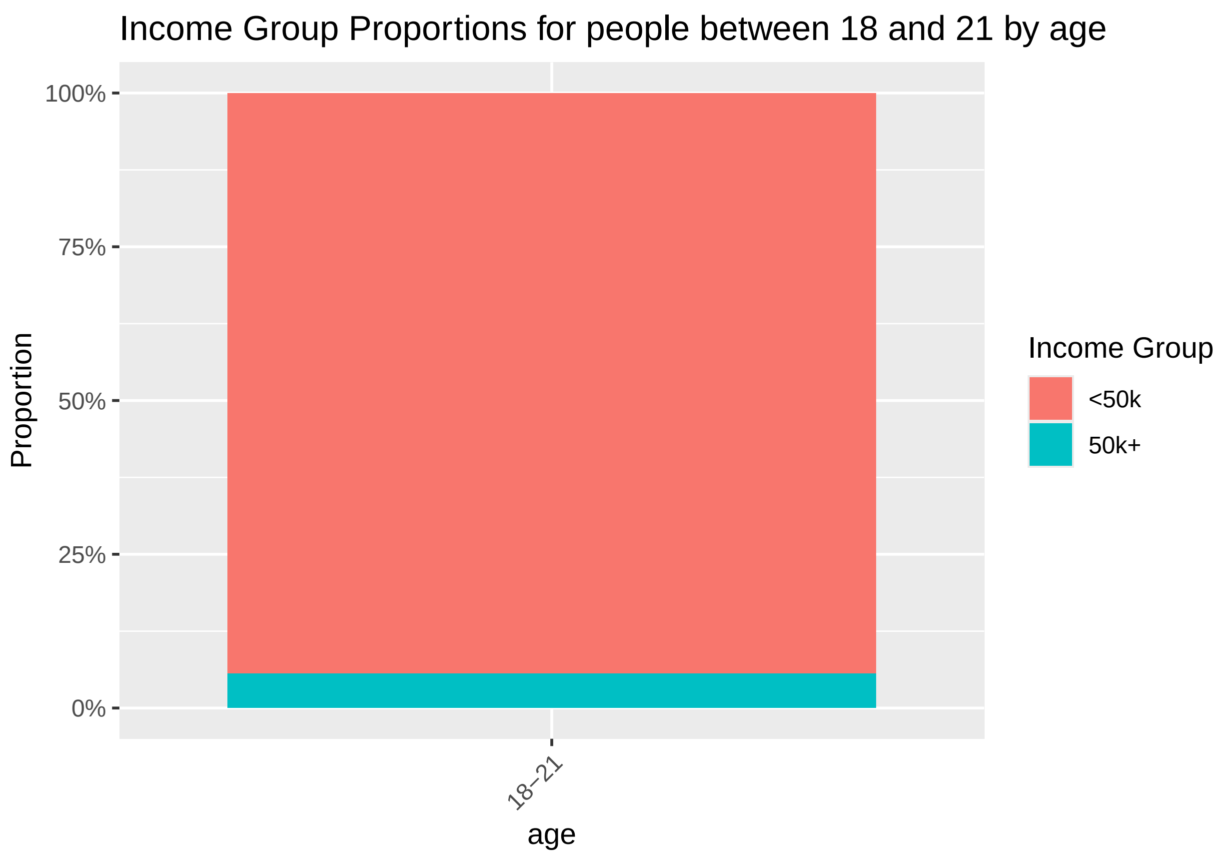
First we thought it might had something to do with the feature: ML_atwork. The easiest way to see if our hypothesis is right, is to disprove it. So we gave everyone that uses ML at work earns less than 50 thousand and is younger than 25 or has less than 3 years of experience a flag to disprove our hypothesis. Now if there was almost nobody that had this new flag then we could conclude that ML at work does in fact have something to do with it.

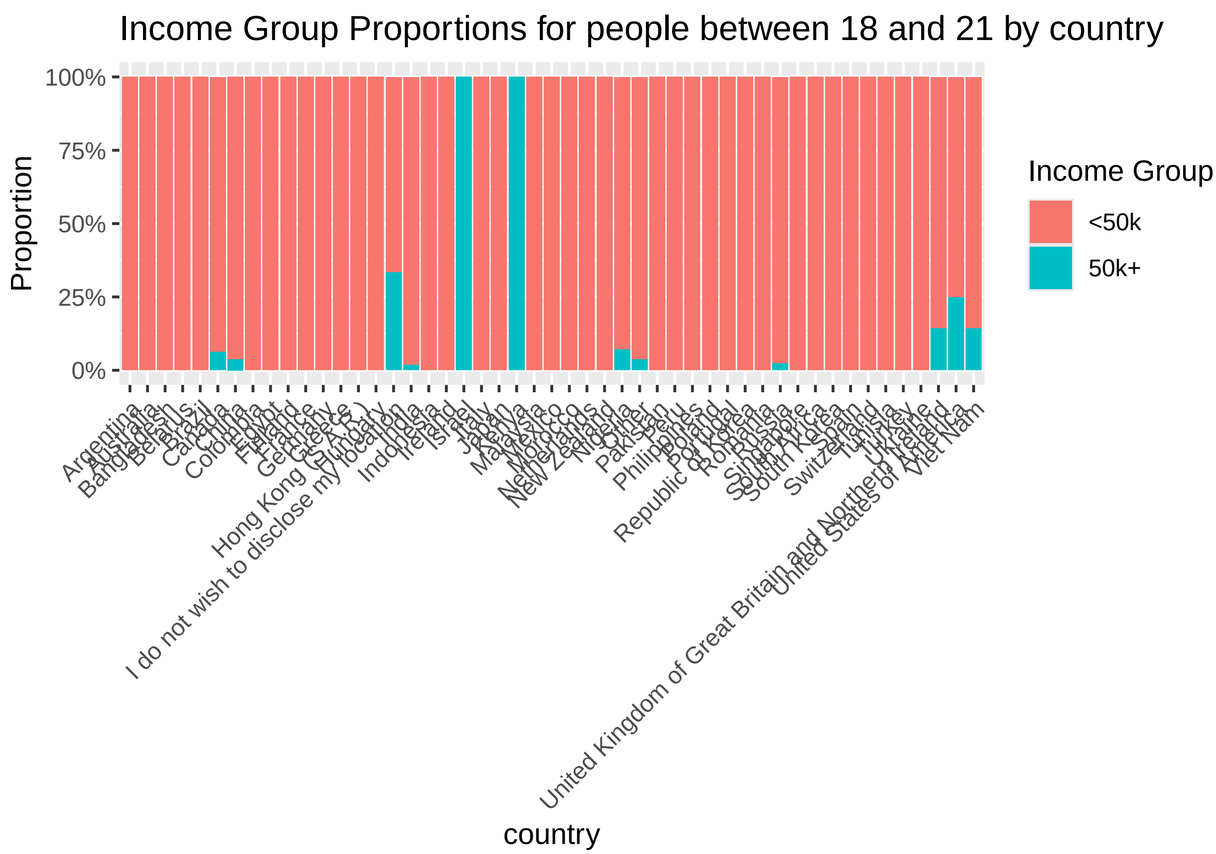
```
df$disporve_outlier_flag <- with(df,
                                 (wage < 50000 & (years_experience %in% c("0-1", "1-2") & age %in% c("1
)
```

Sadly this wasn't the case, as there are a whole lot of people who earn less than 50k a year who also use ML_atwork. So this feature alone didn't cause the high wages for these young/ barely experienced people. Time to do some further analysis. There are 78 features and over 10 thousand rows. With just looking at the data will be hard to find these features that are the cause of these high earners, that's why we will plot some charts to get a clear look of what these high earners do, or in what category they fall. We will plot charts for every categorical feature.

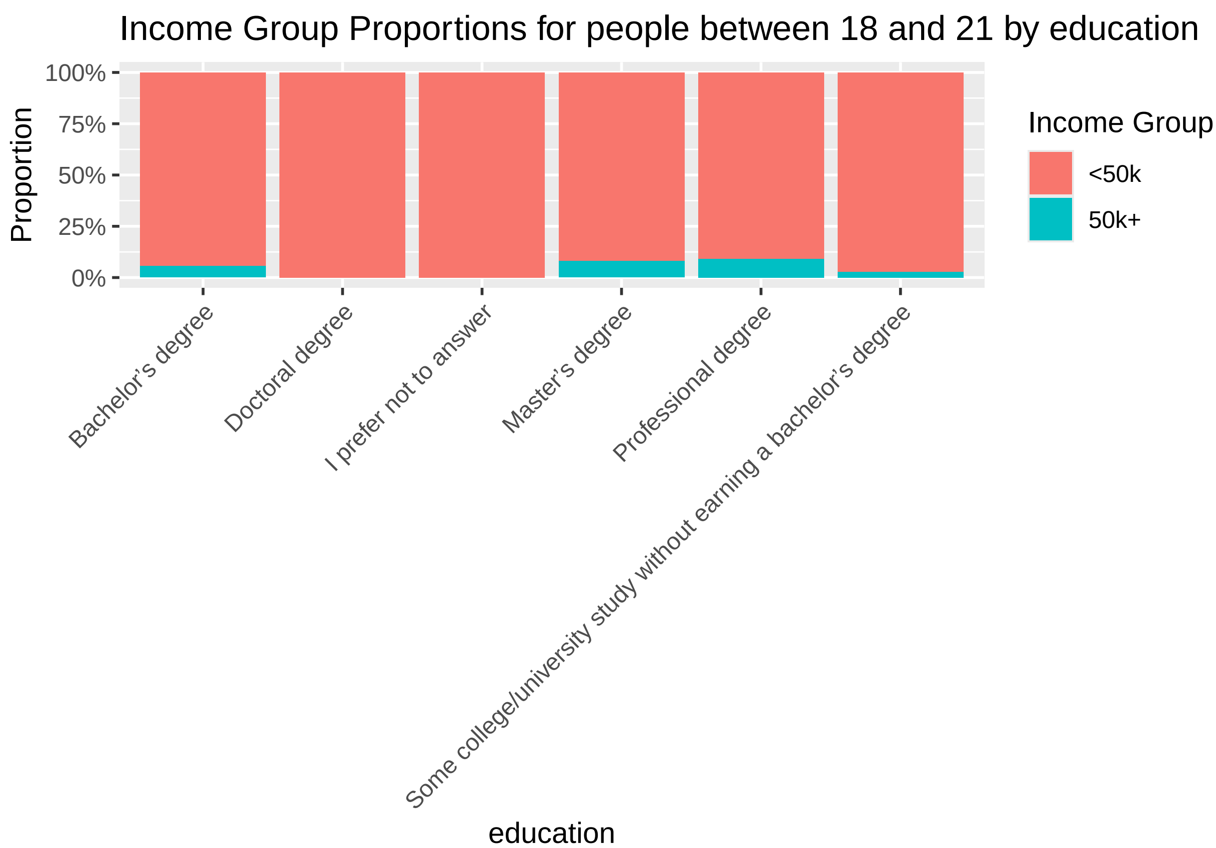
```
df_18_21 <- subset(df, age == "18-21")
df_18_21$income_group <- ifelse(df_18_21$wage >= 50000, "50k+", "<50k")
df_18_21$income_group <- factor(df_18_21$income_group, levels = c("<50k", "50k+"))
features_to_plot <- names(df_18_21)[!(names(df_18_21) %in% c("wage", "income_group", "outlier_flag", "d
cat_features <- features_to_plot[sapply(df_18_21[features_to_plot], function(x) is.character(x) || is.f

# Loop through and plot each categorical feature
for (feature in cat_features) {
  p <- ggplot(df_18_21, aes_string(x = feature, fill = "income_group")) +
    geom_bar(position = "fill") +
    scale_y_continuous(labels = scales::percent_format()) +
    labs(title = paste("Income Group Proportions for people between 18 and 21 by", feature),
         x = feature,
         y = "Proportion",
         fill = "Income Group") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

  print(p)
}
```

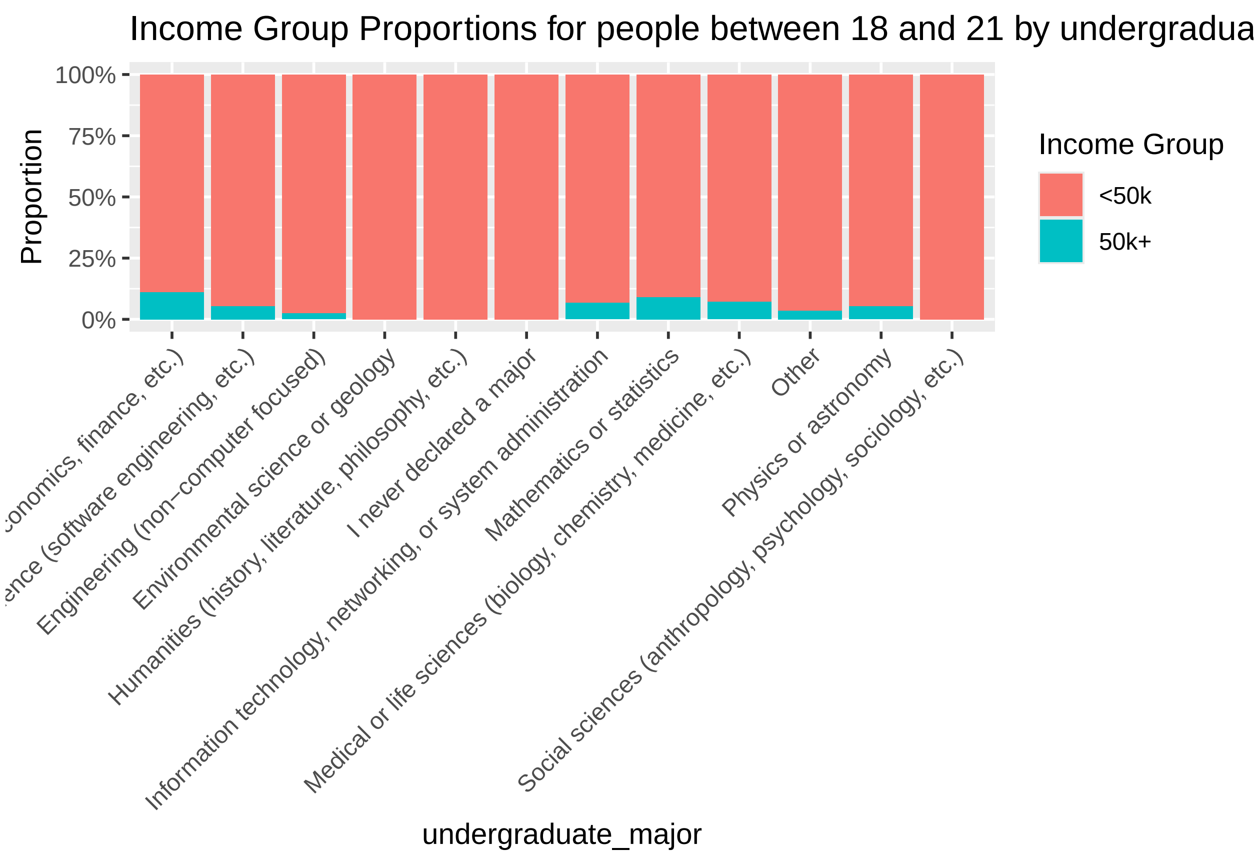

Income Group Proportions for people between 18 and 21 by gender
100%
75%
50%
25%
0%
Proportion
Female
Male
Prefer not to say
Prefer to self–describe
gender
Income Group
<50k
50k+

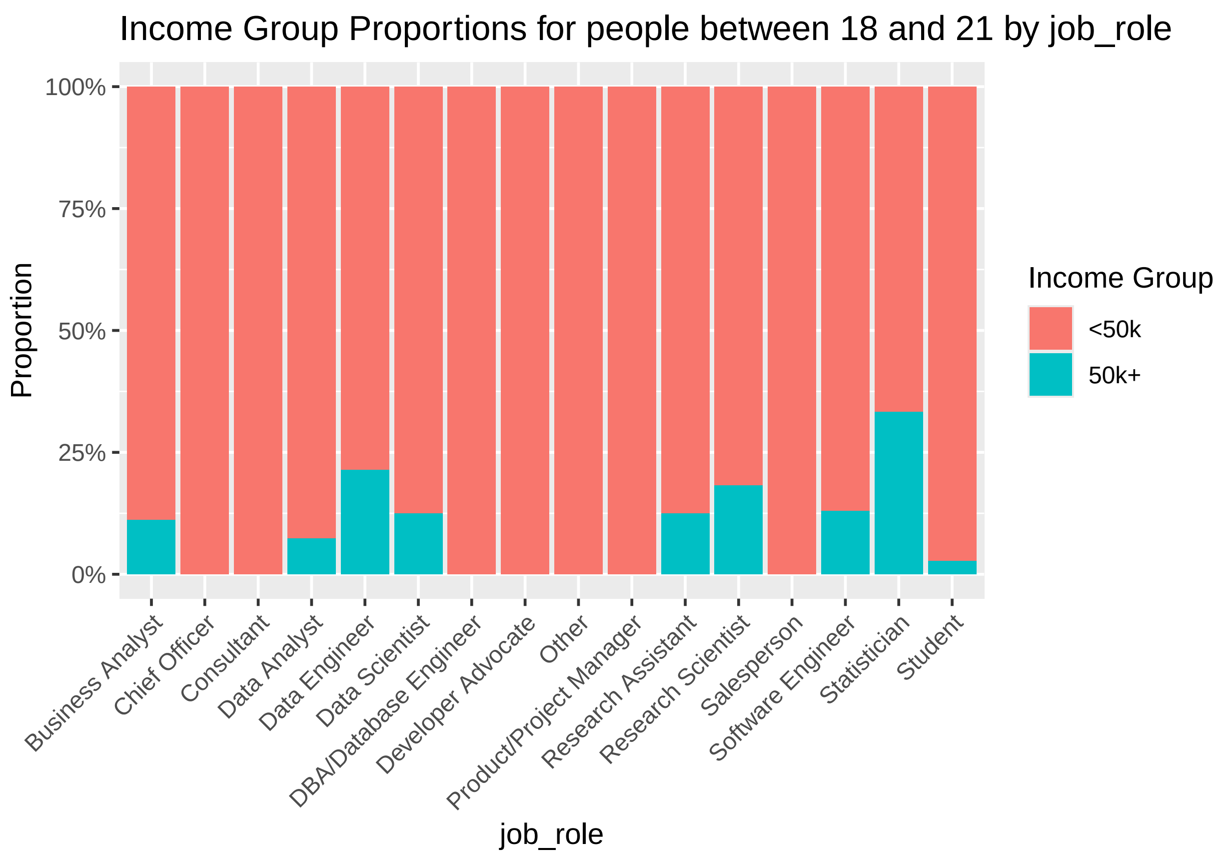

Income Group Proportions for people between 18 and 21 by age
100%
75%
50%
25%
0%
Proportion
18–21
age
Income Group
<50k
50k+

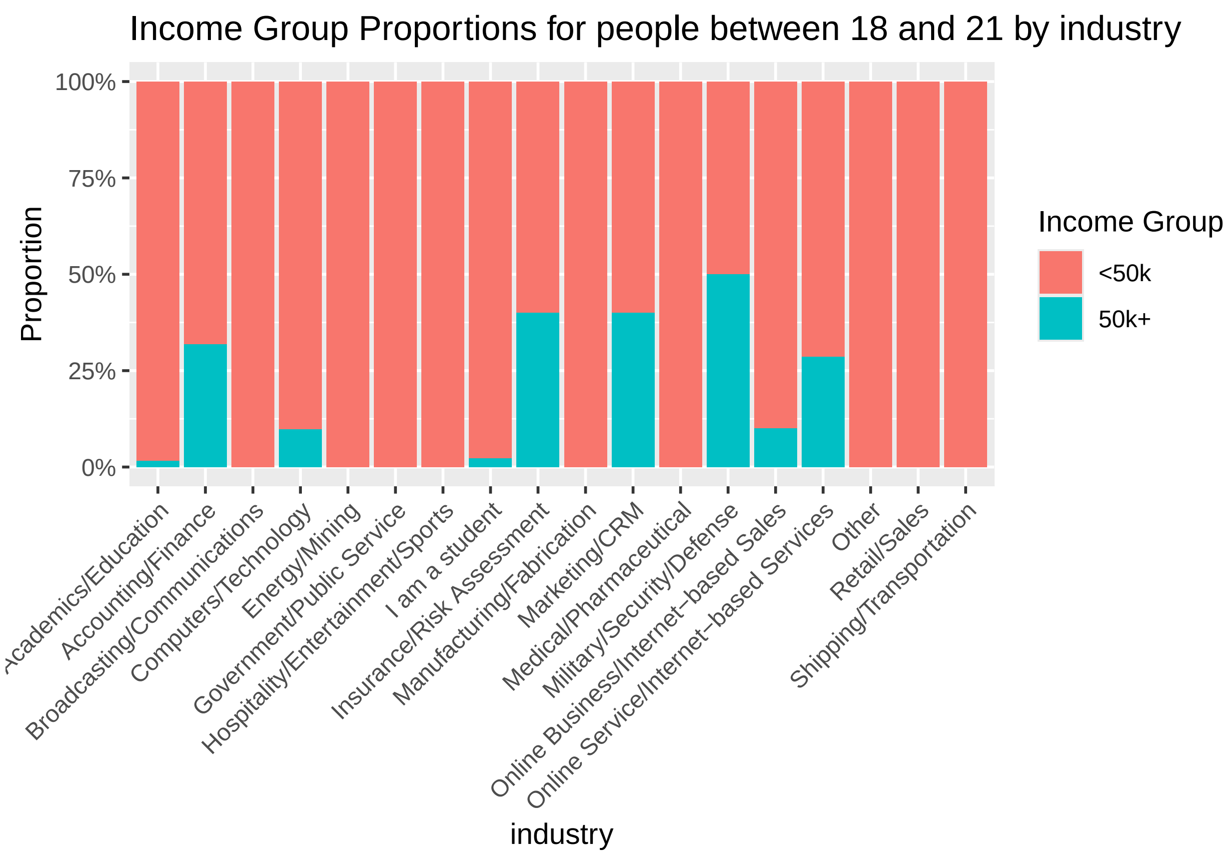


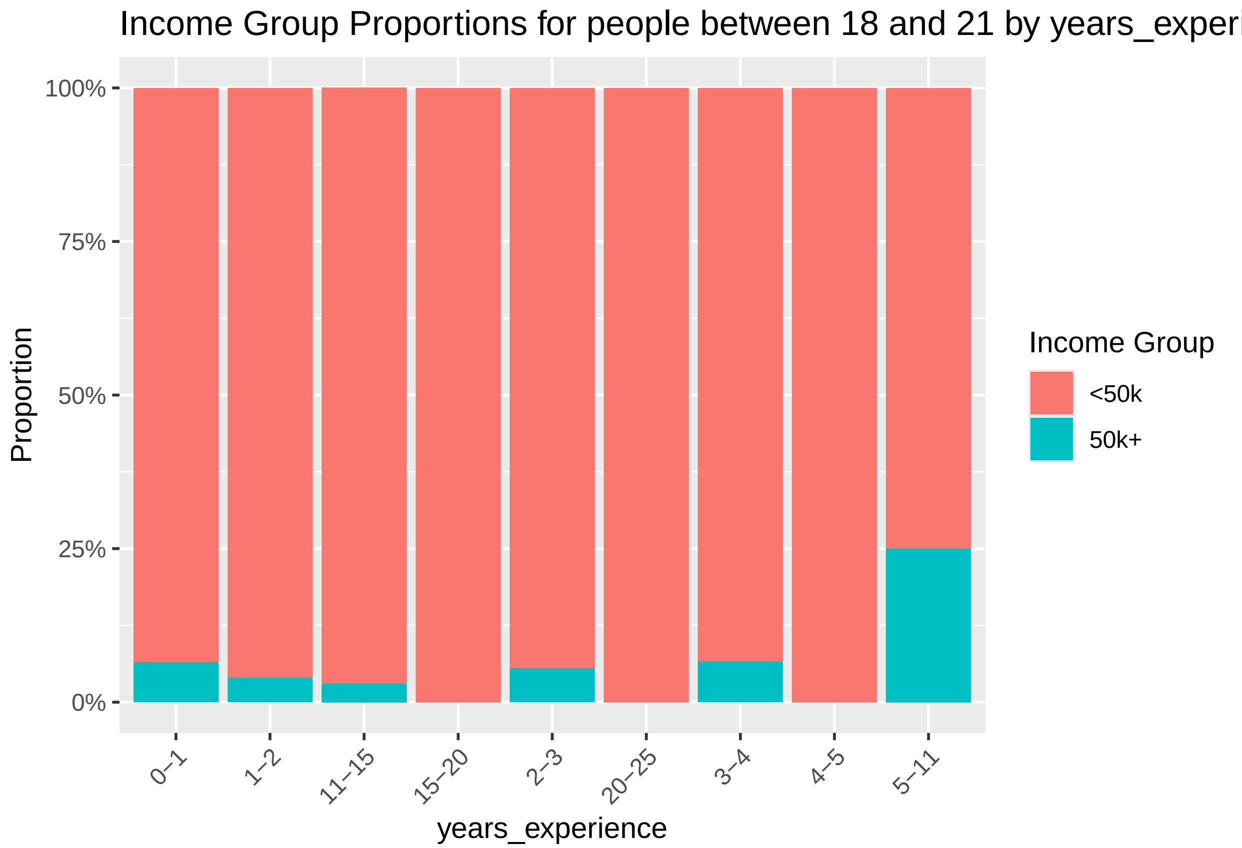
Income Group Proportions for people between 18 and 21 by country
Proportion
100%
75%
50%
25%
0%
country
Income Group
<50k
50k+

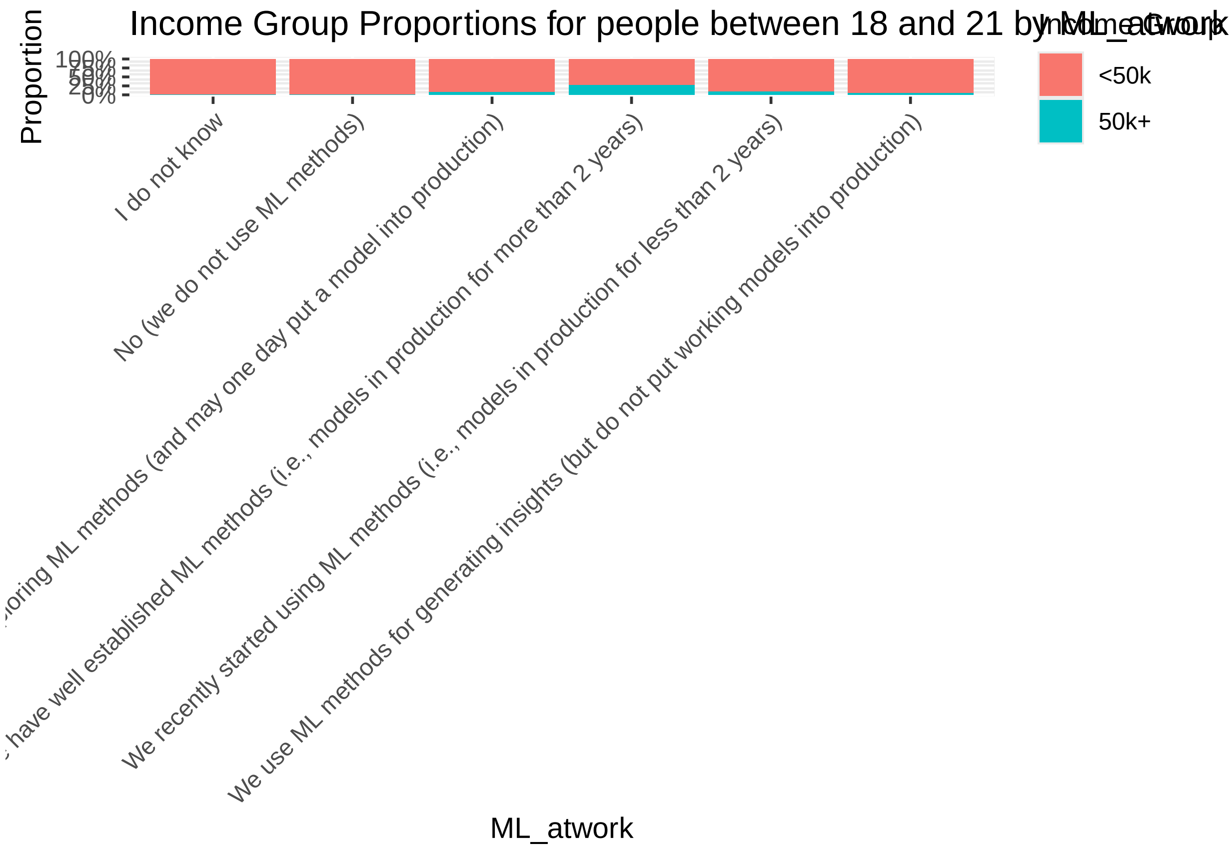

Income Group Proportions for people between 18 and 21 by education
Proportion
100%
75%
50%
25%
0%
Bachelor's degree
Doctoral degree
I prefer not to answer
Master's degree
Professional degree
Some college/university study without earning a bachelor's degree
education
Income Group
<50k
50k+


Income Group Proportions for people between 18 and 21 by undergradua
Proportion
100%
75%
50%
25%
0%
Income Group
<50k
50k+
conomics, finance, etc.)
ence (software engineering, etc.)
Engineering (non−computer focused)
Environmental science or geology
Humanities (history, literature, philosophy, etc.)
I never declared a major
Information technology, networking, or system administration
Mathematics or statistics
Medical or life sciences (biology, chemistry, medicine, etc.)
Other
Physics or astronomy
Social sciences (anthropology, psychology, sociology, etc.)
undergraduate_major

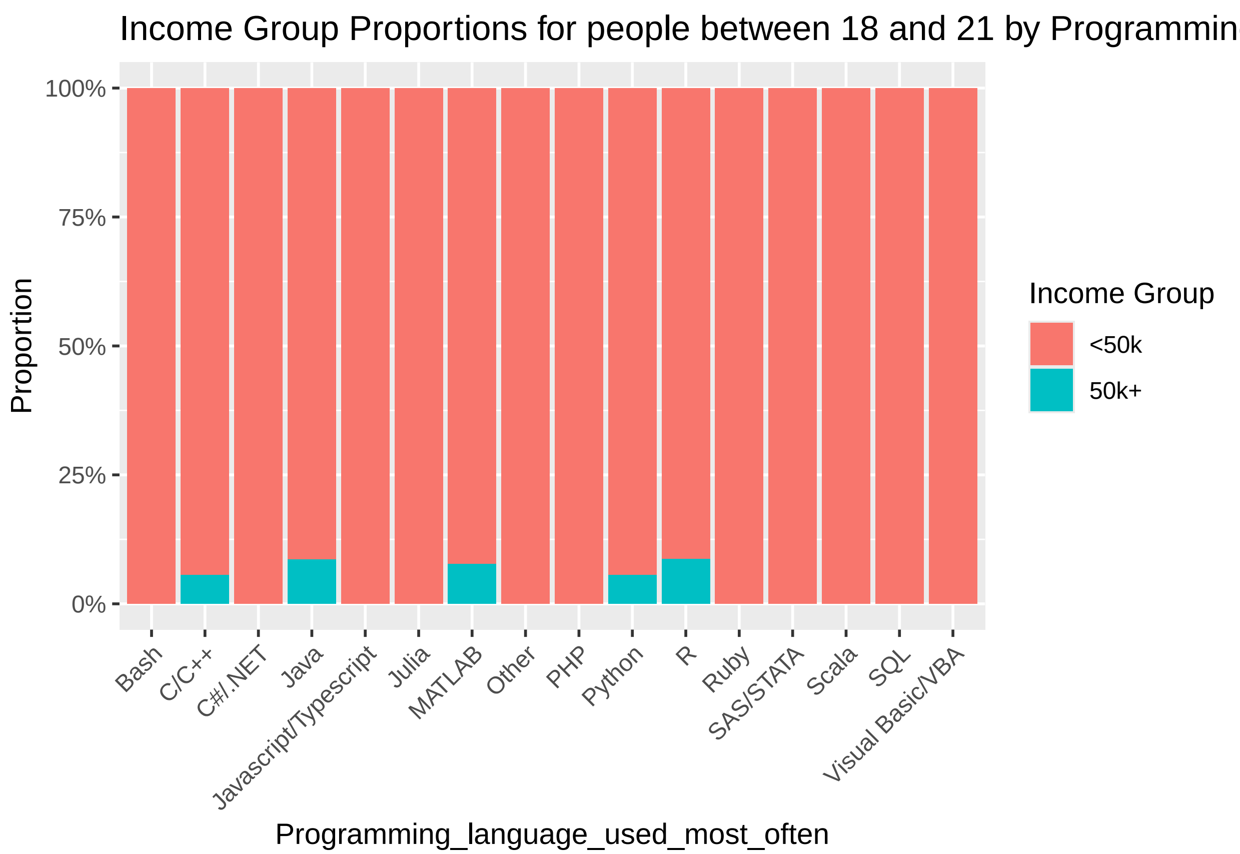

Income Group Proportions for people between 18 and 21 by job_role
100%
75%
50%
25%
0%
Proportion
Business Analyst
Chief Officer
Consultant
Data Analyst
Data Engineer
Data Scientist
DBA/Database Engineer
Developer Advocate
Other
Product/Project Manager
Research Assistant
Research Scientist
Salesperson
Software Engineer
Statistician
Student
job_role
Income Group
<50k
50k+

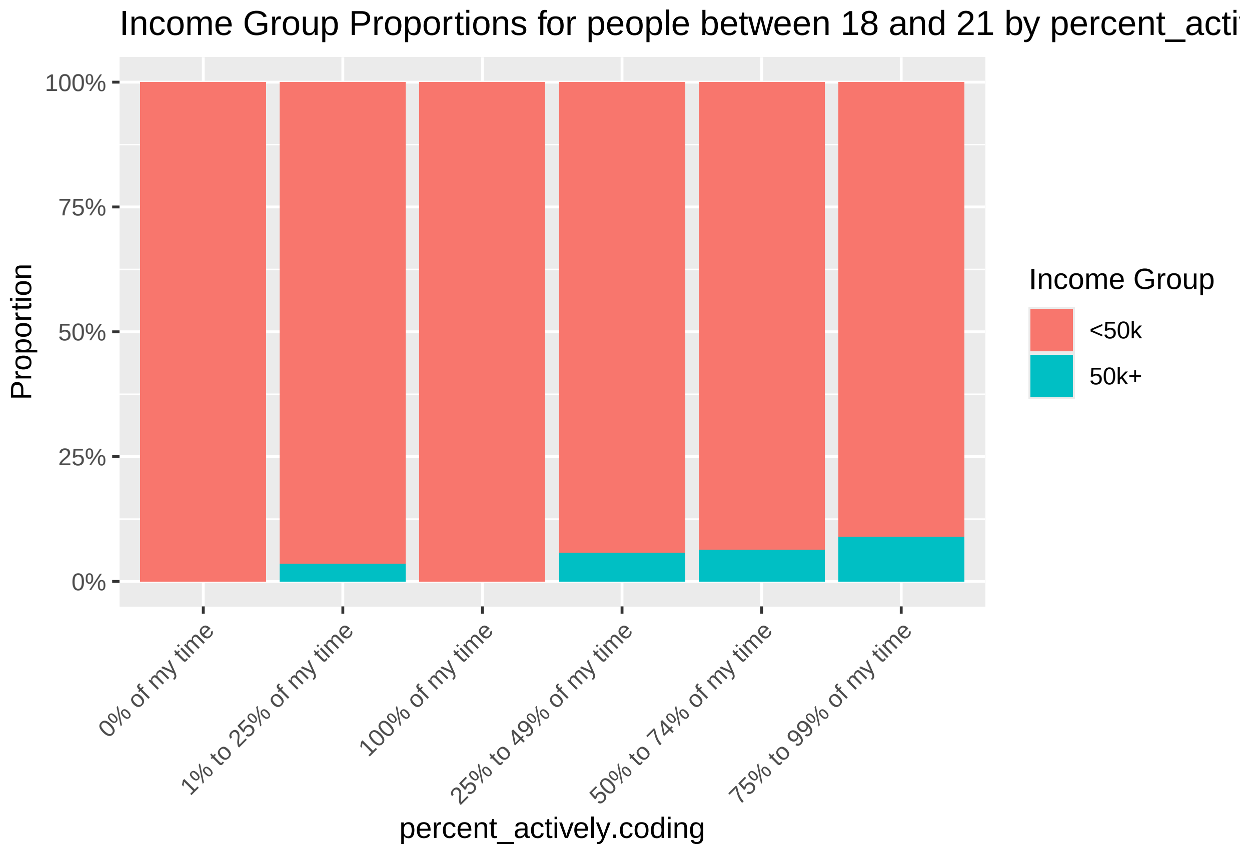

Income Group Proportions for people between 18 and 21 by industry
Proportion
100%
75%
50%
25%
0%
Academics/Education
Accounting/Finance
Broadcasting/Communications
Computers/Technology
Energy/Mining
Government/Public Service
Hospitality/Entertainment/Sports
I am a student
Insurance/Risk Assessment
Manufacturing/Fabrication
Marketing/CRM
Medical/Pharmaceutical
Military/Security/Defense
Online Business/Internet−based Sales
Online Service/Internet−based Services
Other
Retail/Sales
Shipping/Transportation
industry
Income Group
<50k
50k+

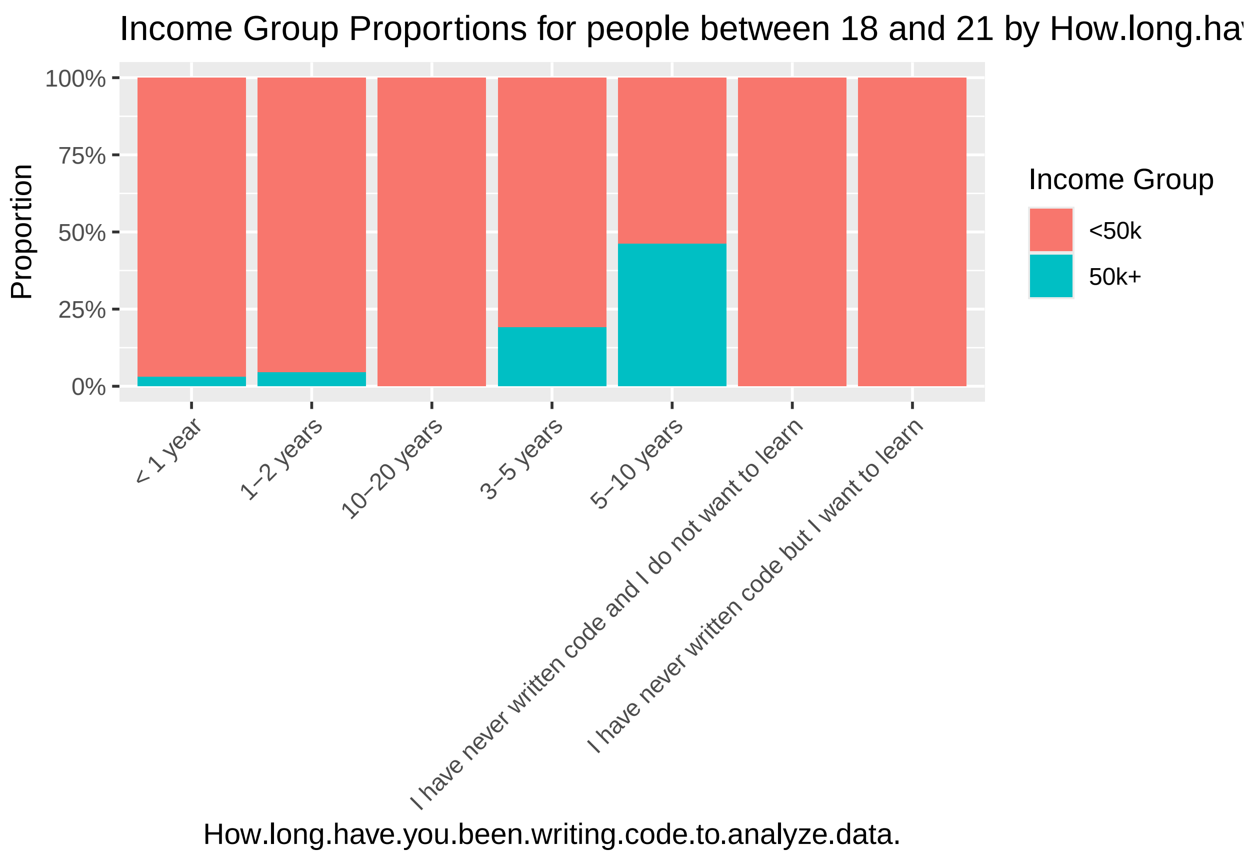

Income Group Proportions for people between 18 and 21 by years_experi
100%
75%
50%
25%
0%
Proportion
0–1
1–2
11–15
15–20
2–3
20–25
3–4
4–5
5–11
years_experience
Income Group
<50k
50k+



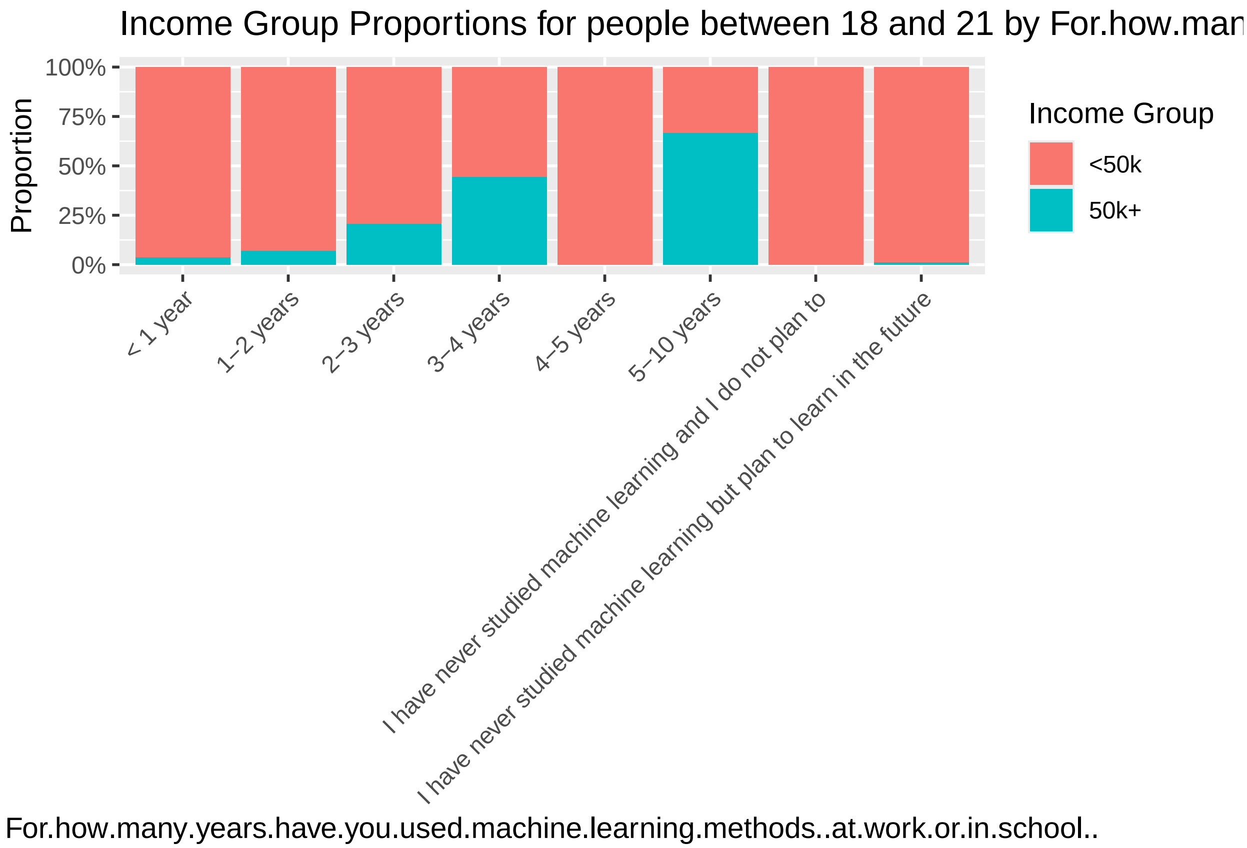
Income Group Proportions for people between 18 and 21 by ML_atwork
Proportion
100%
75%
50%
25%
0%
I do not know
No (we do not use ML methods)
...oring ML methods (and may one day put a model into production)
... have well established ML methods (i.e., models in production for more than 2 years)
We recently started using ML methods (i.e., models in production for less than 2 years)
We use ML methods for generating insights (but do not put working models into production)
ML_atwork
Income Group
<50k
50k+

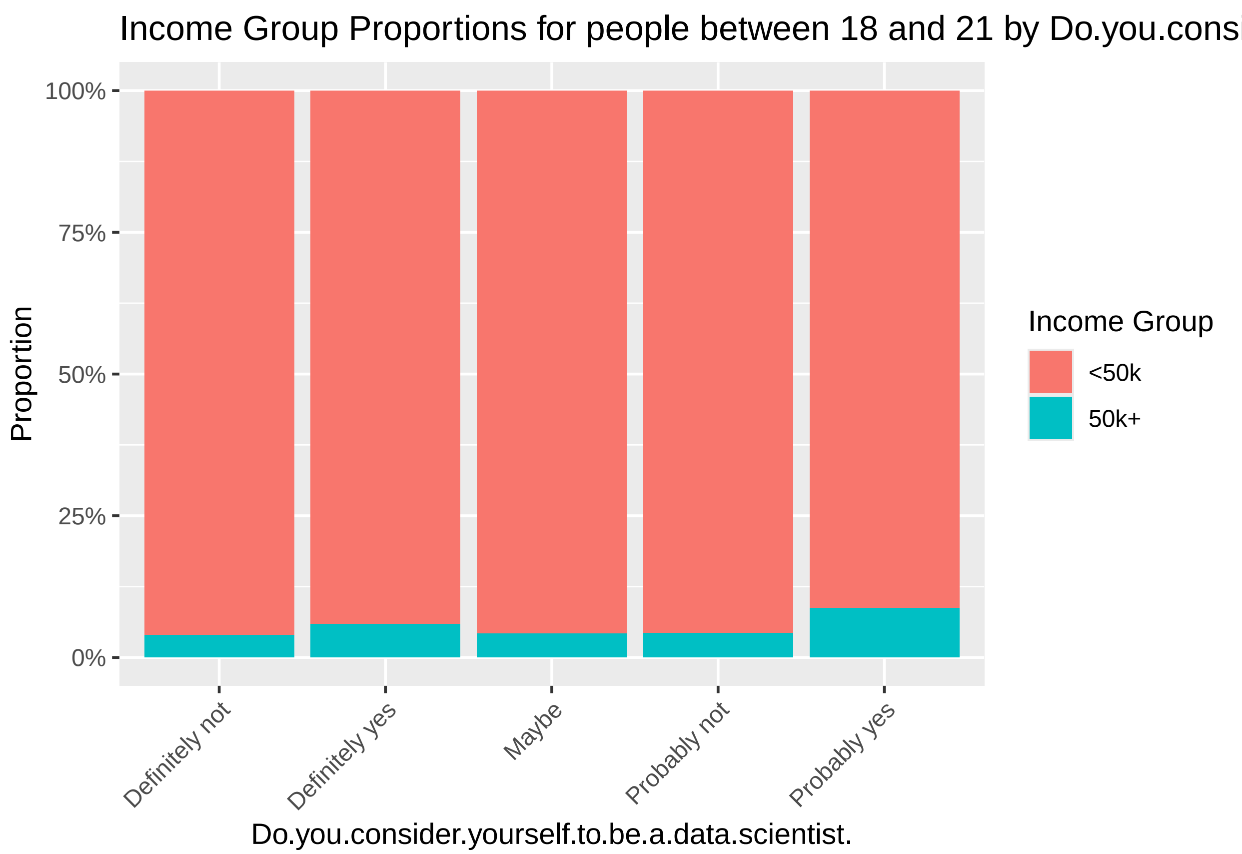

Income Group Proportions for people between 18 and 21 by Programming
100%
75%
50%
25%
0%
Proportion
Bash
C/C++
C#/.NET
Java
Javascript/Typescript
Julia
MATLAB
Other
PHP
Python
R
Ruby
SAS/STATA
Scala
SQL
Visual Basic/VBA
Programming_language_used_most_often
Income Group
<50k
50k+


Income Group Proportions for people between 18 and 21 by percent_acti
Proportion
100%
75%
50%
25%
0%
0% of my time
1% to 25% of my time
100% of my time
25% to 49% of my time
50% to 74% of my time
75% to 99% of my time
percent_actively.coding
Income Group
<50k
50k+



Income Group Proportions for people between 18 and 21 by How.long.ha
100%
75%
50%
25%
0%
Proportion
Income Group
<50k
50k+
< 1 year
1–2 years
10–20 years
3–5 years
5–10 years
I have never written code and I do not want to learn
I have never written code but I want to learn
How.long.have.you.been.writing.code.to.analyze.data.

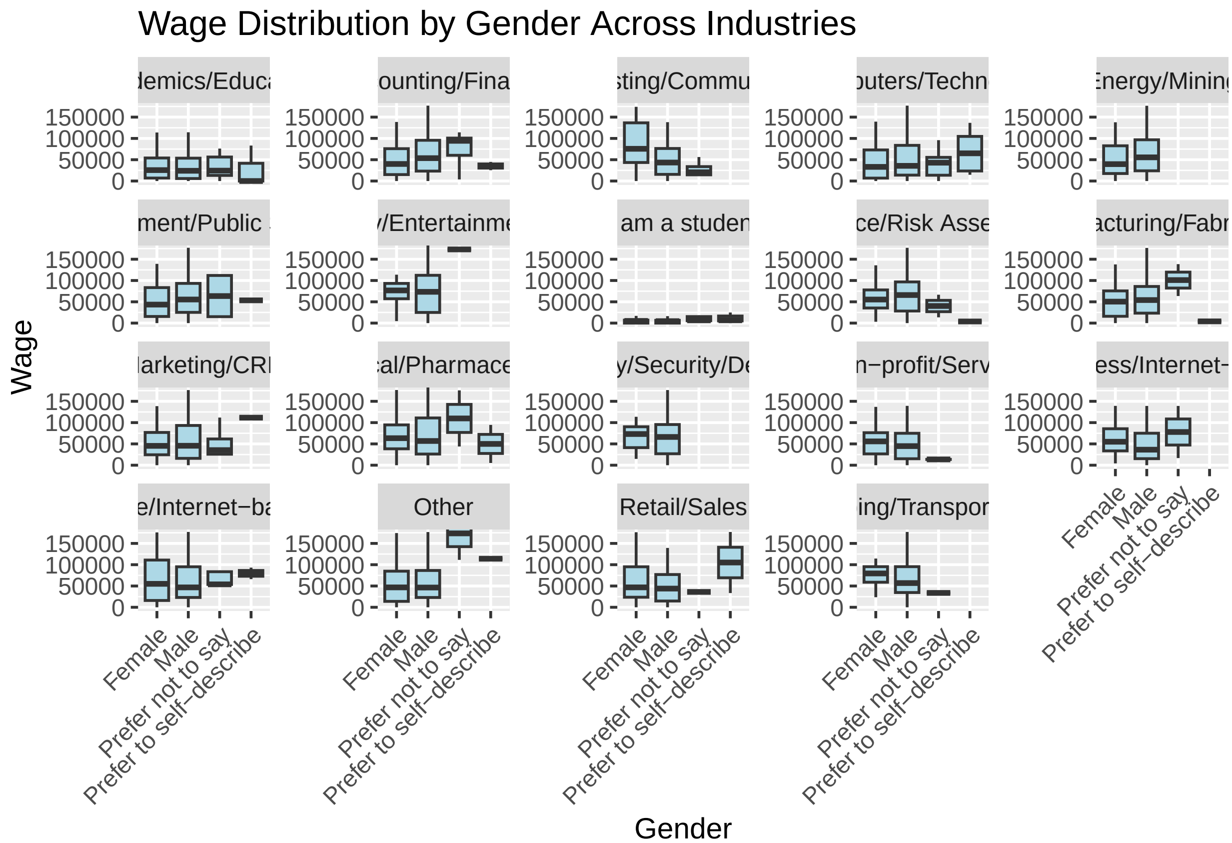

Income Group Proportions for people between 18 and 21 by For.how.man
Proportion
100%
75%
50%
25%
0%
< 1 year
1–2 years
2–3 years
3–4 years
4–5 years
5–10 years
I have never studied machine learning and I do not plan to
I have never studied machine learning but plan to learn in the future
For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school..
Income Group
<50k
50k+



From this we could see some interesting statistics, just check out the plots, “percent actively coding”, “How long have you been writing code to analyze data”, and “for how many years have you used machine learning methods at work or in school”. We can see a clear trend in these categories, so these are some features we got to keep an eye out for. Other graphs like the one “Do you consider yourself to be a data scientist” will not help us at all, as every column has around the same amount of people who earn a lot, making this noise across all categories for that feature.

A normal question is should we consider gender as an important feature? While gender used to heavily affect the wage, does it still to this day and should we include this? Is this ethical?

```
ggplot(df, aes(x = gender, y = wage)) +
  geom_boxplot(fill = "lightblue", outlier.shape = NA) +
  coord_cartesian(ylim = c(0, quantile(df$wage, 0.95))) +  # optional: cap extreme outliers
  labs(
    title = "Wage Distribution by Gender Across Industries",
    y = "Wage",
    x = "Gender"
  ) +
  facet_wrap(~ industry, scales = "free_y") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Gender includes Female, Male, Prefer not to say, and Prefer to self-describe. The difference between the median of all these is very minor, so you might think we could let this variable out. But if we look closer, per industry for example, we see that females on average earn a lot more than males, if we would let the feature gender go and we would try to predict a female or male for the industry broadcasting, it would give a very wrong answer. So even though it might not be ethical, we do have to leave the feature in.

```r
# Numeric correlation with wage
numeric_vars <- sapply(df, is.numeric)
if (sum(numeric_vars) > 1) {
  cor_wage <- cor(df[, numeric_vars], use = "complete.obs")
  print(cor_wage["wage", ])
}
```

```
##                                          Activities_Analyze.and.understand.data.to.influence.produc
##
##                        Activities_Build.and.or.run.a.machine.learning.service.that.operationally.improves.
##
## Activities_Build.and.or.run.the.data.infrastructure.that.my.business.uses.for.storing..analyzing..an
##
##                                              Activities_Build.prototypes.to.explore.applying.machin
##
##                                              Activities_Do.research.that.advances.the.state.of.the.
##
##                                               Activities_None.of.these.activities.are.an.important.
##
##                                                                                                   N
##
```

```
##
##
##                                                                    No
##
##                                                              Notebool
##
##                                                                 Notel
##
##
##
##                                                            cloud_Googl
##
##                                                              cloud_Ama
##
##
##
##
##
##
##
##                                                       cloud_I.have.not.us
##
##
##
##
##
##
##
##
##
##
##
##                                                            Programming
##
##                                                                 Progra
##
##
##
##
##
##
##
##
##
##
##
##                                                                    ML_
##
##                                                                      I
##
##
##
##
##
```

```
##                                                                    M
##
##
##
##
##
##
##
##                                                                   ML
##
##
##
##
##
##                                                                    Vi
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##
##                                              explainability.model_Examine.indivi
##
##                                                   explainability.model_exami
##
##                                                    explainability.model_Exam
##
##                                                 explainability.model_Create.p
##
##                                                           explainabili
##
```

```
##                                                                                                explainabili
##
##                                                     explainability.model_None.I.do.not.use.these.model
##
##
##
```

```r
# Identify all numeric variables (excluding the target variable "wage")
numeric_vars <- setdiff(names(df)[sapply(df, is.numeric)], "wage")

# Loop through each numeric variable and run linear regression
for (var in numeric_vars) {
  formula <- as.formula(paste("wage ~", var))
  lm_result <- lm(formula, data = df)

  cat("\nLinear regression of wage on", var, ":\n")
  print(summary(lm_result))
}
```

```
##
## Linear regression of wage on Activities_Analyze.and.understand.data.to.influence.product.or.business
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63812 -37473 -17281  22153 510646
##
## Coefficients:
##                                                                          Estimate
## (Intercept)                                                               40359.9
## Activities_Analyze.and.understand.data.to.influence.product.or.business.decisions  23451.6
##                                                                          Std. Error
## (Intercept)                                                                   863.9
## Activities_Analyze.and.understand.data.to.influence.product.or.business.decisions    1174.6
##                                                                          t value
## (Intercept)                                                                46.72
## Activities_Analyze.and.understand.data.to.influence.product.or.business.decisions   19.97
##                                                                          Pr(>|t|)
## (Intercept)                                                                <2e-16
## Activities_Analyze.and.understand.data.to.influence.product.or.business.decisions   <2e-16
##
## (Intercept)                                                              ***
## Activities_Analyze.and.understand.data.to.influence.product.or.business.decisions ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60850 on 10807 degrees of freedom
## Multiple R-squared:  0.03558,    Adjusted R-squared:  0.03549
## F-statistic: 398.7 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on Activities_Build.and.or.run.a.machine.learning.service.that.operational
```

```
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67854 -41225 -20218  26117 505465
##
## Coefficients:
##
## (Intercept)
## Activities_Build.and.or.run.a.machine.learning.service.that.operationally.improves.my.product.or.worl
##
## (Intercept)
## Activities_Build.and.or.run.a.machine.learning.service.that.operationally.improves.my.product.or.worl
##
## (Intercept)
## Activities_Build.and.or.run.a.machine.learning.service.that.operationally.improves.my.product.or.worl
##
## (Intercept)
## Activities_Build.and.or.run.a.machine.learning.service.that.operationally.improves.my.product.or.worl
##
## (Intercept)
## Activities_Build.and.or.run.a.machine.learning.service.that.operationally.improves.my.product.or.worl
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61150 on 10807 degrees of freedom
## Multiple R-squared:  0.02599,    Adjusted R-squared:  0.0259
## F-statistic: 288.4 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on Activities_Build.and.or.run.the.data.infrastructure.that.my.business.use
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67030 -41455 -20445  20282 504756
##
## Coefficients:
##
## (Intercept)
## Activities_Build.and.or.run.the.data.infrastructure.that.my.business.uses.for.storing..analyzing..and
##
## (Intercept)
## Activities_Build.and.or.run.the.data.infrastructure.that.my.business.uses.for.storing..analyzing..and
##
## (Intercept)
## Activities_Build.and.or.run.the.data.infrastructure.that.my.business.uses.for.storing..analyzing..and
##
## (Intercept)
## Activities_Build.and.or.run.the.data.infrastructure.that.my.business.uses.for.storing..analyzing..and
```

```
##
## (Intercept)
## Activities_Build.and.or.run.the.data.infrastructure.that.my.business.uses.for.storing..analyzing..and
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61240 on 10807 degrees of freedom
## Multiple R-squared:  0.02316,    Adjusted R-squared:  0.02307
## F-statistic: 256.2 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on Activities_Build.prototypes.to.explore.applying.machine.learning.to.new
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68475 -37336 -16600  23679 510031
##
## Coefficients:
##                                                                           Estimate
## (Intercept)                                                                41571.3
## Activities_Build.prototypes.to.explore.applying.machine.learning.to.new.areas  26903.3
##                                                                           Std. Error
## (Intercept)                                                                    768.7
## Activities_Build.prototypes.to.explore.applying.machine.learning.to.new.areas     1176.9
##                                                                           t value
## (Intercept)                                                                 54.08
## Activities_Build.prototypes.to.explore.applying.machine.learning.to.new.areas   22.86
##                                                                           Pr(>|t|)
## (Intercept)                                                                 <2e-16
## Activities_Build.prototypes.to.explore.applying.machine.learning.to.new.areas   <2e-16
##
## (Intercept)                                                               ***
## Activities_Build.prototypes.to.explore.applying.machine.learning.to.new.areas ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60520 on 10807 degrees of freedom
## Multiple R-squared:  0.04612,    Adjusted R-squared:  0.04603
## F-statistic: 522.5 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on Activities_Do.research.that.advances.the.state.of.the.art.of.machine.lea
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -55371 -45495 -18464  23202 499366
##
## Coefficients:
```

```
##                                                                              Estimate
## (Intercept)                                                                  52236.3
## Activities_Do.research.that.advances.the.state.of.the.art.of.machine.learning   3134.6
##                                                                             Std. Error
## (Intercept)                                                                     692.2
## Activities_Do.research.that.advances.the.state.of.the.art.of.machine.learning     1360.2
##                                                                             t value
## (Intercept)                                                                  75.467
## Activities_Do.research.that.advances.the.state.of.the.art.of.machine.learning   2.304
##                                                                             Pr(>|t|)
## (Intercept)                                                                   <2e-16
## Activities_Do.research.that.advances.the.state.of.the.art.of.machine.learning   0.0212
##
## (Intercept)                                                                 ***
## Activities_Do.research.that.advances.the.state.of.the.art.of.machine.learning *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61950 on 10807 degrees of freedom
## Multiple R-squared:  0.0004912,  Adjusted R-squared:  0.0003987
## F-statistic: 5.311 on 1 and 10807 DF,  p-value: 0.02121
##
##
## Linear regression of wage on Activities_None.of.these.activities.are.an.important.part.of.my.role.at
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56475 -41456 -19588  20629 495299
##
## Coefficients:
##                                                                            Estimate
## (Intercept)                                                                 56474.7
## Activities_None.of.these.activities.are.an.important.part.of.my.role.at.work -22245.6
##                                                                            Std. Error
## (Intercept)                                                                    642.5
## Activities_None.of.these.activities.are.an.important.part.of.my.role.at.work     1637.1
##                                                                            t value
## (Intercept)                                                                 87.89
## Activities_None.of.these.activities.are.an.important.part.of.my.role.at.work  -13.59
##                                                                            Pr(>|t|)
## (Intercept)                                                                  <2e-16
## Activities_None.of.these.activities.are.an.important.part.of.my.role.at.work   <2e-16
##
## (Intercept)                                                                ***
## Activities_None.of.these.activities.are.an.important.part.of.my.role.at.work ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61440 on 10807 degrees of freedom
## Multiple R-squared:  0.0168, Adjusted R-squared:  0.01671
## F-statistic: 184.6 on 1 and 10807 DF,  p-value: < 2.2e-16
```

```
##
##
## Linear regression of wage on Notebooks_Kaggle.Kernels :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -54750 -43227 -19106  22121 501798
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               54749.8      729.5  75.052  < 2e-16 ***
## Notebooks_Kaggle.Kernels  -5102.6     1263.2  -4.039 5.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61920 on 10807 degrees of freedom
## Multiple R-squared:  0.001508,   Adjusted R-squared:  0.001415
## F-statistic: 16.32 on 1 and 10807 DF,  p-value: 5.394e-05
##
##
## Linear regression of wage on Notebooks_Google.Colab :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53899 -43265 -18605  22322 501584
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             53899.2      663.8  81.203  < 2e-16 ***
## Notebooks_Google.Colab  -4379.0     1505.5  -2.909  0.00364 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61940 on 10807 degrees of freedom
## Multiple R-squared:  0.0007822,  Adjusted R-squared:  0.0006898
## F-statistic:  8.46 on 1 and 10807 DF,  p-value: 0.003637
##
##
## Linear regression of wage on Notebooks_Azure.Notebook :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57643 -45876 -18125  22926 499093
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               52680.2      619.3  85.071   <2e-16 ***
## Notebooks_Azure.Notebook   4963.2     2274.8   2.182   0.0291 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61950 on 10807 degrees of freedom
## Multiple R-squared:  0.0004403,  Adjusted R-squared:  0.0003478
## F-statistic:  4.76 on 1 and 10807 DF,  p-value: 0.02914
##
##
## Linear regression of wage on Notebooks_Google.Cloud.Datalab :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -60021 -45763 -17933  23160 499277
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     52496.6      618.8  84.836  < 2e-16 ***
## Notebooks_Google.Cloud.Datalab   7524.7     2286.0   3.292 0.000999 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61930 on 10807 degrees of freedom
## Multiple R-squared:  0.001002,   Adjusted R-squared:  0.0009091
## F-statistic: 10.83 on 1 and 10807 DF,  p-value: 0.0009994
##
##
## Linear regression of wage on Notebooks_JupyterHub.Binder :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56451 -45026 -18002  23653 499703
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  51741.9      700.7  73.843  < 2e-16 ***
## Notebooks_JupyterHub.Binder   4708.8     1330.5   3.539 0.000403 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61930 on 10807 degrees of freedom
## Multiple R-squared:  0.001158,   Adjusted R-squared:  0.001065
## F-statistic: 12.53 on 1 and 10807 DF,  p-value: 0.000403
##
##
## Linear regression of wage on Notebooks_None :
```

```
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53131 -46237 -18267  22639 498776
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      52997.3      756.7  70.040   <2e-16 ***
## Notebooks_None     133.6     1228.1   0.109    0.913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61960 on 10807 degrees of freedom
## Multiple R-squared:  1.094e-06,  Adjusted R-squared:  -9.144e-05
## F-statistic: 0.01183 on 1 and 10807 DF,  p-value: 0.9134
##
##
## Linear regression of wage on cloud_Google.Cloud.Platform..GCP. :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -60556 -44327 -16634  24479 501582
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        50191.5      698.3  71.877  < 2e-16 ***
## cloud_Google.Cloud.Platform..GCP.  10364.5     1330.1   7.792  7.2e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61790 on 10807 degrees of freedom
## Multiple R-squared:  0.005587,   Adjusted R-squared:  0.005495
## F-statistic: 60.72 on 1 and 10807 DF,  p-value: 7.198e-15
##
##
## Linear regression of wage on cloud_Amazon.Web.Services..AWS. :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68635 -35896 -15596  24490 511215
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      39790.9      788.6   50.45   <2e-16 ***
## cloud_Amazon.Web.Services..AWS.  28843.7     1163.3   24.80   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60270 on 10807 degrees of freedom
## Multiple R-squared:  0.05383,    Adjusted R-squared:  0.05374
## F-statistic: 614.8 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on cloud_Microsoft.Azure :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63821 -43800 -17319  24604 501996
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            49777.8      677.3   73.49   <2e-16 ***
## cloud_Microsoft.Azure  14043.7     1403.6   10.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61680 on 10807 degrees of freedom
## Multiple R-squared:  0.009178,   Adjusted R-squared:  0.009086
## F-statistic: 100.1 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on cloud_IBM.Cloud :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -60834 -45771 -17902  23076 499300
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      52473.3      617.2  85.013  < 2e-16 ***
## cloud_IBM.Cloud   8360.3     2354.3   3.551 0.000385 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61930 on 10807 degrees of freedom
## Multiple R-squared:  0.001166,   Adjusted R-squared:  0.001073
## F-statistic: 12.61 on 1 and 10807 DF,  p-value: 0.0003852
##
##
## Linear regression of wage on cloud_Alibaba.Cloud :
##
## Call:
## lm(formula = formula, data = df)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53319 -43271 -18366  22436 507833
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)            53319        604   88.28  < 2e-16 ***
## cloud_Alibaba.Cloud   -10048       3681   -2.73  0.00635 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61940 on 10807 degrees of freedom
## Multiple R-squared:  0.000689,   Adjusted R-squared:  0.0005965
## F-statistic: 7.451 on 1 and 10807 DF,  p-value: 0.006351
##
##
## Linear regression of wage on cloud_I.have.not.used.any.cloud.providers :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -60622 -37021 -16659  24140 511439
##
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                60622.4      711.7   85.18   <2e-16
## cloud_I.have.not.used.any.cloud.providers -23601.0     1256.3  -18.79   <2e-16
##
## (Intercept)                               ***
## cloud_I.have.not.used.any.cloud.providers ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60980 on 10807 degrees of freedom
## Multiple R-squared:  0.03162,    Adjusted R-squared:  0.03153
## F-statistic: 352.9 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on Programming_Python :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53139 -46228 -18268  22641 498738
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         53139.2     1744.2  30.465   <2e-16 ***
## Programming_Python   -103.2     1856.0  -0.056    0.956
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61960 on 10807 degrees of freedom
## Multiple R-squared:  2.863e-07,  Adjusted R-squared:  -9.225e-05
## F-statistic: 0.003094 on 1 and 10807 DF,  p-value: 0.9556
##
##
## Linear regression of wage on Programming_R :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -58682 -43323 -21863  24989 502818
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    48955.8      780.7   62.70  < 2e-16 ***
## Programming_R   9725.8     1203.6    8.08 7.14e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61780 on 10807 degrees of freedom
## Multiple R-squared:  0.006006,   Adjusted R-squared:  0.005914
## F-statistic: 65.29 on 1 and 10807 DF,  p-value: 7.14e-16
##
##
## Linear regression of wage on Programming_SQL :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59372 -41310 -19288  24098 506386
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      45387.2      880.7   51.54   <2e-16 ***
## Programming_SQL  13985.0     1189.9   11.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61570 on 10807 degrees of freedom
## Multiple R-squared:  0.01262,    Adjusted R-squared:  0.01253
## F-statistic: 138.1 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on Programming_Bash :
##
## Call:
## lm(formula = formula, data = df)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -69011 -43382 -16147  25171 502541
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       49232.8      658.1   74.81   <2e-16 ***
## Programming_Bash  19778.5     1498.5   13.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61470 on 10807 degrees of freedom
## Multiple R-squared:  0.01587,    Adjusted R-squared:  0.01577
## F-statistic: 174.2 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on Programming_Java :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -55061 -41300 -19677  21347 504621
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       55061.0      680.8  80.874  < 2e-16 ***
## Programming_Java  -8519.4     1400.6  -6.083 1.22e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61860 on 10807 degrees of freedom
## Multiple R-squared:  0.003412,   Adjusted R-squared:  0.00332
## F-statistic:    37 on 1 and 10807 DF,  p-value: 1.222e-09
##
##
## Linear regression of wage on Programming_Javascript.Typescript :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -55040 -45756 -18365  22932 499261
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         52512.2      671.3  78.226   <2e-16 ***
## Programming_Javascript.Typescript   2528.0     1458.1   1.734    0.083 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 61960 on 10807 degrees of freedom
## Multiple R-squared:  0.0002781,  Adjusted R-squared:  0.0001856
## F-statistic: 3.006 on 1 and 10807 DF,  p-value: 0.08299
##
##
## Linear regression of wage on Programming_Visual.Basic.VBA :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -60680 -45656 -17793  23196 499426
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   52347.2      622.3  84.116  < 2e-16 ***
## Programming_Visual.Basic.VBA   8332.7     2146.0   3.883 0.000104 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61920 on 10807 degrees of freedom
## Multiple R-squared:  0.001393,   Adjusted R-squared:  0.001301
## F-statistic: 15.08 on 1 and 10807 DF,  p-value: 0.0001038
##
##
## Linear regression of wage on Programming_C.C.. :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57138 -40753 -20513  22652 511020
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        57137.7      683.5   83.60   <2e-16 ***
## Programming_C.C.. -16384.4     1368.1  -11.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61560 on 10807 degrees of freedom
## Multiple R-squared:  0.0131, Adjusted R-squared:  0.01301
## F-statistic: 143.4 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on Programming_MATLAB :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -55558 -40544 -19441  20987 511760
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         55557.6      645.4  86.087   <2e-16 ***
## Programming_MATLAB -16214.1     1640.4  -9.884   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61690 on 10807 degrees of freedom
## Multiple R-squared:  0.008959,   Adjusted R-squared:  0.008868
## F-statistic:  97.7 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on Programming_Scala :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75676 -45130 -17554  23615 500117
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        51656.9      611.5  84.473   <2e-16 ***
## Programming_Scala  24019.2     2541.1   9.452   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61710 on 10807 degrees of freedom
## Multiple R-squared:  0.0082, Adjusted R-squared:  0.008108
## F-statistic: 89.35 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on Programming_Julia :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -76217 -45970 -18044  22840 499080
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        52693.3      599.9  87.837  < 2e-16 ***
## Programming_Julia  23523.7     4885.1   4.815 1.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61900 on 10807 degrees of freedom
## Multiple R-squared:  0.002141,   Adjusted R-squared:  0.002049
## F-statistic: 23.19 on 1 and 10807 DF,  p-value: 1.489e-06
```

```
##
##
## Linear regression of wage on Programming_SAS.STATA :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -62749 -45707 -17889  23019 499455
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            52318.4      617.4  84.736  < 2e-16 ***
## Programming_SAS.STATA  10430.9     2334.6   4.468 7.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61910 on 10807 degrees of freedom
## Multiple R-squared:  0.001844,   Adjusted R-squared:  0.001751
## F-statistic: 19.96 on 1 and 10807 DF,  p-value: 7.98e-06
##
##
## Linear regression of wage on ML_framework_Scikit.Learn :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -55675 -41771 -20130  21227 504915
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  46858       1090  43.010  < 2e-16 ***
## ML_framework_Scikit.Learn     8817       1300   6.781 1.26e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61830 on 10807 degrees of freedom
## Multiple R-squared:  0.004237,   Adjusted R-squared:  0.004145
## F-statistic: 45.98 on 1 and 10807 DF,  p-value: 1.257e-11
##
##
## Linear regression of wage on ML_framework_TensorFlow :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56150 -42765 -20310  25096 502839
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               48934.8      907.5  53.925   <2e-16 ***
## ML_framework_TensorFlow   7215.2     1201.9   6.003    2e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61860 on 10807 degrees of freedom
## Multiple R-squared:  0.003324,   Adjusted R-squared:  0.003232
## F-statistic: 36.04 on 1 and 10807 DF,  p-value: 1.996e-09
##
##
## Linear regression of wage on ML_framework_Keras :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -57528 -43242 -21228  25497 502669
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          49105.1      815.3  60.226  < 2e-16 ***
## ML_framework_Keras   8422.7     1191.7   7.068 1.67e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61820 on 10807 degrees of freedom
## Multiple R-squared:  0.004601,   Adjusted R-squared:  0.004509
## F-statistic: 49.96 on 1 and 10807 DF,  p-value: 1.669e-12
##
##
## Linear regression of wage on ML_framework_PyTorch :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -58833 -44953 -17540  24048 500322
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            51451.2      672.4  76.516  < 2e-16 ***
## ML_framework_PyTorch   7382.1     1445.8   5.106 3.35e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61890 on 10807 degrees of freedom
## Multiple R-squared:  0.002406,   Adjusted R-squared:  0.002314
## F-statistic: 26.07 on 1 and 10807 DF,  p-value: 3.35e-07
##
##
## Linear regression of wage on ML_framework_Spark.MLlib :
```

```
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75026 -43413 -16275  24674 502256
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               49517.5      635.6   77.91   <2e-16 ***
## ML_framework_Spark.MLlib  25508.8     1708.4   14.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61330 on 10807 degrees of freedom
## Multiple R-squared:  0.02021,    Adjusted R-squared:  0.02012
## F-statistic:   223 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on ML_framework_H20 :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -78876 -44211 -16821  24166 501232
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         50542        619   81.66   <2e-16 ***
## ML_framework_H20    28334       2081   13.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61440 on 10807 degrees of freedom
## Multiple R-squared:  0.01686,    Adjusted R-squared:  0.01677
## F-statistic: 185.3 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on ML_framework_Caret :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67451 -44234 -17024  23962 501175
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         50598.7      641.7   78.85   <2e-16 ***
## ML_framework_Caret  16852.2     1683.2   10.01   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61680 on 10807 degrees of freedom
## Multiple R-squared:  0.00919,    Adjusted R-squared:  0.009098
## F-statistic: 100.2 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on ML_framework_Xgboost :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63877 -42078 -19962  25354 504102
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           47672.0      723.5   65.89   <2e-16 ***
## ML_framework_Xgboost  16204.5     1256.2   12.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61490 on 10807 degrees of freedom
## Multiple R-squared:  0.01516,    Adjusted R-squared:  0.01507
## F-statistic: 166.4 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on ML_framework_randomForest :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59837 -43721 -16390  24658 502353
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                49420.7      735.9  67.160   <2e-16 ***
## ML_framework_randomForest  10416.5     1247.0   8.353   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61760 on 10807 degrees of freedom
## Multiple R-squared:  0.006415,   Adjusted R-squared:  0.006323
## F-statistic: 69.78 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on ML_framework_None :
##
## Call:
## lm(formula = formula, data = df)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -54888 -40717 -19270  21713 496886
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        54887.9      632.5   86.78   <2e-16 ***
## ML_framework_None -15598.4     1841.6   -8.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61760 on 10807 degrees of freedom
## Multiple R-squared:  0.006594,   Adjusted R-squared:  0.006503
## F-statistic: 71.74 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on Visualization_ggplot2 :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59635 -42396 -20450  24603 504487
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            47115.0      817.5   57.63   <2e-16 ***
## Visualization_ggplot2  12520.5     1187.5   10.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61650 on 10807 degrees of freedom
## Multiple R-squared:  0.01018,    Adjusted R-squared:  0.01009
## F-statistic: 111.2 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on Visualization_Matplotlib :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -55570 -45451 -18172  23167 499569
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 55570       1190  46.682   <2e-16 ***
## Visualization_Matplotlib    -3365       1375  -2.447   0.0144 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 61950 on 10807 degrees of freedom
## Multiple R-squared:  0.0005539,	Adjusted R-squared:  0.0004614 
## F-statistic: 5.989 on 1 and 10807 DF,  p-value: 0.01441
## 
## 
## Linear regression of wage on Visualization_Altair :
## 
## Call:
## lm(formula = formula, data = df)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -80608 -45910 -18024  22930 499116 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           52657.5      599.4  87.856  < 2e-16 ***
## Visualization_Altair  27950.9     5071.0   5.512 3.63e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 61880 on 10807 degrees of freedom
## Multiple R-squared:  0.002803,	Adjusted R-squared:  0.002711 
## F-statistic: 30.38 on 1 and 10807 DF,  p-value: 3.631e-08
## 
## 
## Linear regression of wage on Visualization_Shiny :
## 
## Call:
## lm(formula = formula, data = df)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -72084 -43005 -18201  24272 502840 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          48933.6      650.5   75.22   <2e-16 ***
## Visualization_Shiny  23150.6     1543.1   15.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 61330 on 10807 degrees of freedom
## Multiple R-squared:  0.0204,	Adjusted R-squared:  0.02031 
## F-statistic: 225.1 on 1 and 10807 DF,  p-value: < 2.2e-16
## 
## 
## Linear regression of wage on Visualization_Plotly :
## 
## Call:
## lm(formula = formula, data = df)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
```

```
## -59018 -43977 -16608  24657 501846
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           49928.0      733.6  68.055  < 2e-16 ***
## Visualization_Plotly   9089.9     1252.2   7.259 4.17e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61810 on 10807 degrees of freedom
## Multiple R-squared:  0.004852,   Adjusted R-squared:  0.00476
## F-statistic: 52.69 on 1 and 10807 DF,  p-value: 4.169e-13
##
##
## Linear regression of wage on Visualization_None :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53881 -40531 -18667  22219 497893
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         53881.1      618.8   87.08  < 2e-16 ***
## Visualization_None -11214.0     2270.1   -4.94 7.94e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61890 on 10807 degrees of freedom
## Multiple R-squared:  0.002253,   Adjusted R-squared:  0.002161
## F-statistic:  24.4 on 1 and 10807 DF,  p-value: 7.938e-07
##
##
## Linear regression of wage on data_Categorical.Data :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59685 -42175 -20177  24623 504416
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            46865.1      823.9   56.88   <2e-16 ***
## data_Categorical.Data  12820.0     1186.4   10.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61630 on 10807 degrees of freedom
## Multiple R-squared:  0.01069,    Adjusted R-squared:  0.0106
## F-statistic: 116.8 on 1 and 10807 DF,  p-value: < 2.2e-16
```

```
##
##
## Linear regression of wage on data_Genetic.Data :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56117 -46046 -18251  22787 498925
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          52849        615  85.939   <2e-16 ***
## data_Genetic.Data     3268       2492   1.311     0.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61960 on 10807 degrees of freedom
## Multiple R-squared:  0.000159,   Adjusted R-squared:  6.649e-05
## F-statistic: 1.719 on 1 and 10807 DF,  p-value: 0.1899
##
##
## Linear regression of wage on data_Geospatial.Data :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -69820 -44037 -16650  24267 501165
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           50280.3      639.4   78.64   <2e-16 ***
## data_Geospatial.Data  19540.1     1698.9   11.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61590 on 10807 degrees of freedom
## Multiple R-squared:  0.01209,    Adjusted R-squared:  0.012
## F-statistic: 132.3 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on data_Image.Data :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -54476 -43067 -18974  22125 502249
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        54476        706  77.156  < 2e-16 ***
## data_Image.Data    -4951       1315  -3.765 0.000167 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61920 on 10807 degrees of freedom
## Multiple R-squared:  0.00131,    Adjusted R-squared:  0.001218
## F-statistic: 14.18 on 1 and 10807 DF,  p-value: 0.0001671
##
##
## Linear regression of wage on data_Numerical.Data :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56573 -41937 -20406  26273 504332
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            46950        982  47.809  < 2e-16 ***
## data_Numerical.Data     9623       1234   7.801 6.72e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61790 on 10807 degrees of freedom
## Multiple R-squared:  0.005599,   Adjusted R-squared:  0.005507
## F-statistic: 60.85 on 1 and 10807 DF,  p-value: 6.725e-15
##
##
## Linear regression of wage on data_Sensor.Data :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63963 -44719 -17450  23837 500695
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        51078.8      645.7  79.104  < 2e-16 ***
## data_Sensor.Data   12884.3     1651.7   7.801 6.72e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61790 on 10807 degrees of freedom
## Multiple R-squared:  0.005599,   Adjusted R-squared:  0.005507
## F-statistic: 60.85 on 1 and 10807 DF,  p-value: 6.724e-15
##
##
## Linear regression of wage on data_Tabular.Data :
```

```
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -60815 -41870 -19926  24046 504881
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)           46892        793   59.13   <2e-16 ***
## data_Tabular.Data    13923       1193   11.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61580 on 10807 degrees of freedom
## Multiple R-squared:  0.01246,    Adjusted R-squared:  0.01236
## F-statistic: 136.3 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on data_text.Data :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -56261 -43201 -19973  24568 501795
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      49807.3      843.6  59.044  < 2e-16 ***
## data_text.Data   6453.3     1190.4   5.421 6.05e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61880 on 10807 degrees of freedom
## Multiple R-squared:  0.002712,   Adjusted R-squared:  0.00262
## F-statistic: 29.39 on 1 and 10807 DF,  p-value: 6.048e-08
##
##
## Linear regression of wage on data_Time.Series.Data :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -62688 -39573 -18566  22094 507099
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             44503.3      809.7   54.96   <2e-16 ***
## data_Time.Series.Data  18184.6     1181.3   15.39   <2e-16 ***
```

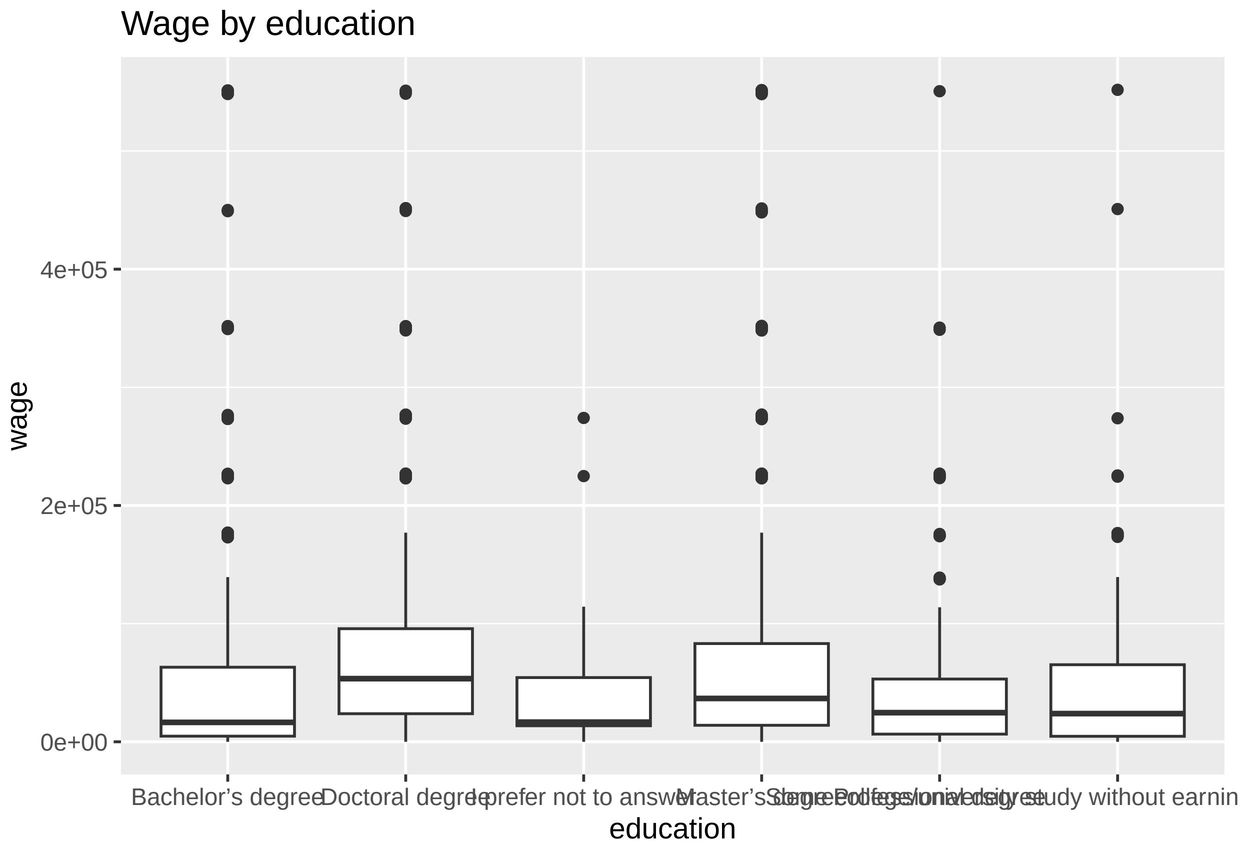
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61300 on 10807 degrees of freedom
## Multiple R-squared:  0.02146,    Adjusted R-squared:  0.02137
## F-statistic:   237 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on data_Video.Data :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53767 -46183 -18294  22680 498786
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      52987.2      620.7  85.373   <2e-16 ***
## data_Video.Data    780.1     2223.8   0.351    0.726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61960 on 10807 degrees of freedom
## Multiple R-squared:  1.139e-05,  Adjusted R-squared:  -8.114e-05
## F-statistic: 0.1231 on 1 and 10807 DF,  p-value: 0.7257
##
##
## Linear regression of wage on explainability.model_Examine.individual.model.coefficients :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -68372 -42724 -21696  25282 503219
##
## Coefficients:
##                                                            Estimate Std. Error
## (Intercept)                                                 48554.3      671.7
## explainability.model_Examine.individual.model.coefficients  19817.6     1410.5
##                                                            t value Pr(>|t|)
## (Intercept)                                                  72.29   <2e-16 ***
## explainability.model_Examine.individual.model.coefficients   14.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61410 on 10807 degrees of freedom
## Multiple R-squared:  0.01794,    Adjusted R-squared:  0.01785
## F-statistic: 197.4 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on explainability.model_examine.feature.correlations :
```

```
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63277 -42156 -19319  23376 503753
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                       47849.8      726.8   65.83
## explainability.model_examine.feature.correlations  15427.6     1252.2   12.32
##                                                  Pr(>|t|)
## (Intercept)                                        <2e-16 ***
## explainability.model_examine.feature.correlations   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61530 on 10807 degrees of freedom
## Multiple R-squared:  0.01385,    Adjusted R-squared:  0.01376
## F-statistic: 151.8 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on explainability.model_Examine.feature.importances :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65769 -40478 -19912  20944 506090
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                      45512.3      742.6   61.28
## explainability.model_Examine.feature.importances  20257.0     1217.6   16.64
##                                                 Pr(>|t|)
## (Intercept)                                       <2e-16 ***
## explainability.model_Examine.feature.importances   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61190 on 10807 degrees of freedom
## Multiple R-squared:  0.02497,    Adjusted R-squared:  0.02488
## F-statistic: 276.8 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on explainability.model_Create.partial.dependence.plots :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

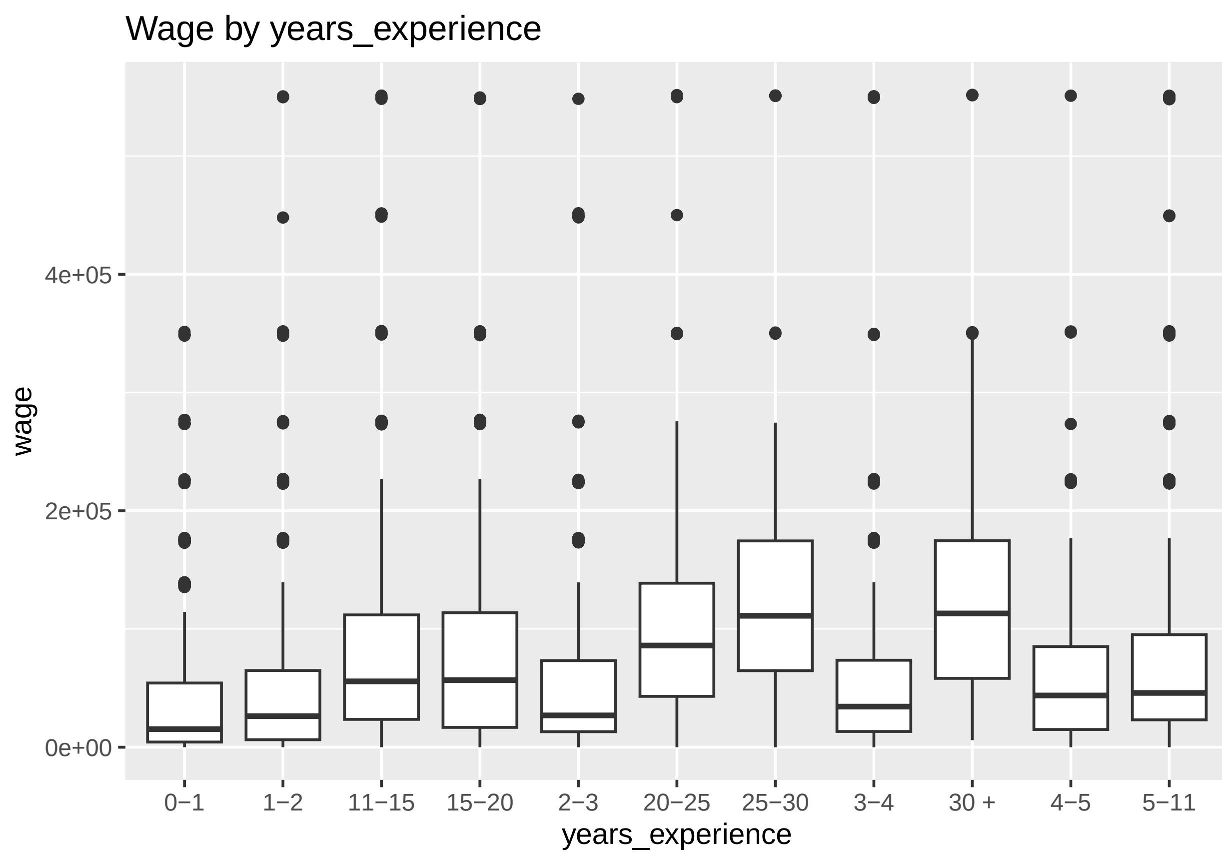
```
## -66710 -44729 -17288  24040 500574
##
## Coefficients:
##                                                     Estimate Std. Error
## (Intercept)                                          51199.8      632.9
## explainability.model_Create.partial.dependence.plots  15510.0     1833.6
##                                                     t value Pr(>|t|)
## (Intercept)                                          80.892   <2e-16 ***
## explainability.model_Create.partial.dependence.plots   8.459   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61760 on 10807 degrees of freedom
## Multiple R-squared:  0.006577,	Adjusted R-squared:  0.006486
## F-statistic: 71.55 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on explainability.model_LIME.functions :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -80302 -44786 -17471  23816 500464
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          51309.9      610.9   83.99   <2e-16 ***
## explainability.model_LIME.functions  28991.8     2495.0   11.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61580 on 10807 degrees of freedom
## Multiple R-squared:  0.01234,	Adjusted R-squared:  0.01225
## F-statistic:   135 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on explainability.model_SHAP.functions :
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79670 -45208 -17518  23563 500028
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          51745.9      607.6  85.160   <2e-16 ***
## explainability.model_SHAP.functions  27924.5     2814.0   9.923   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

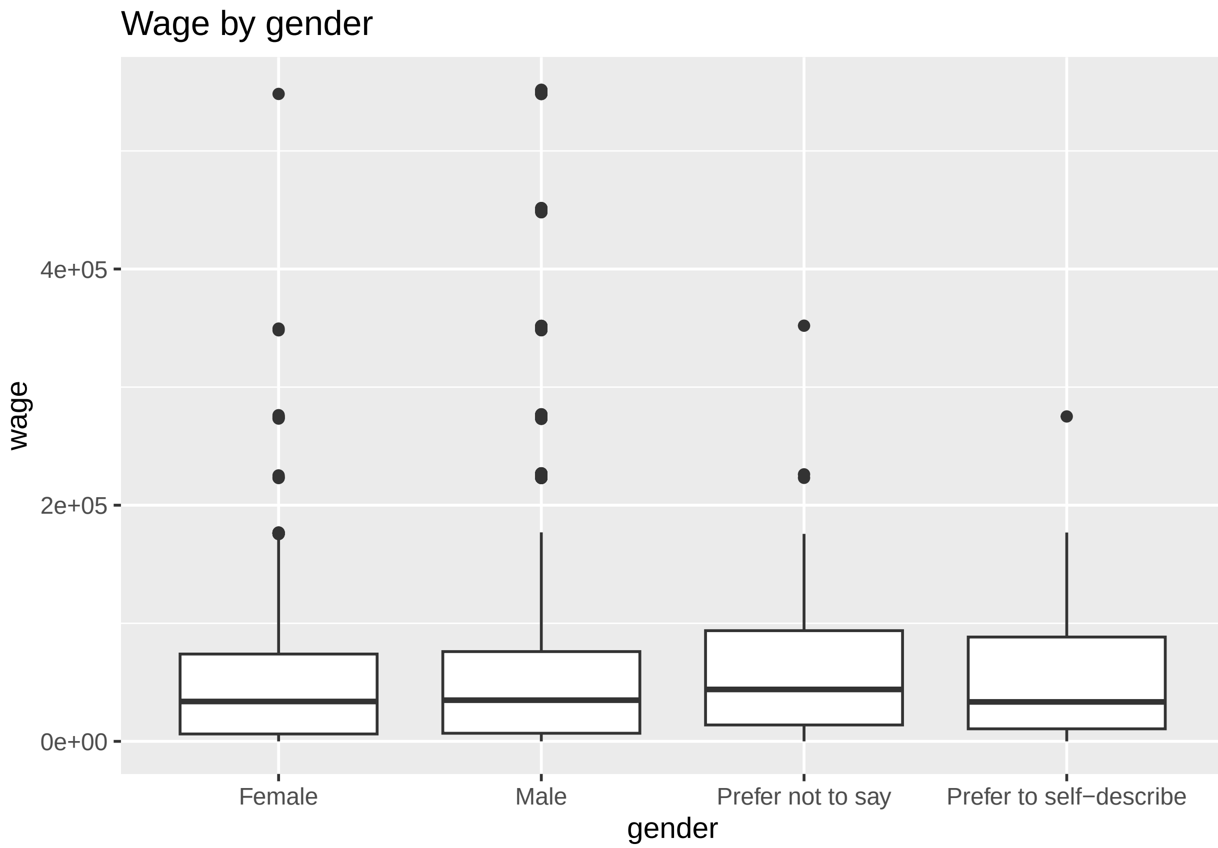
```
## Residual standard error: 61680 on 10807 degrees of freedom
## Multiple R-squared:  0.00903,	Adjusted R-squared:  0.008938
## F-statistic: 98.48 on 1 and 10807 DF,  p-value: < 2.2e-16
##
##
## Linear regression of wage on explainability.model_None.I.do.not.use.these.model.explanation.techniqu
##
## Call:
## lm(formula = formula, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -54312 -40772 -18794  22159 497462
##
## Coefficients:
##                                                                     Estimate
## (Intercept)                                                          54311.6
## explainability.model_None.I.do.not.use.these.model.explanation.techniques -12507.2
##                                                                     Std. Error
## (Intercept)                                                              627.4
## explainability.model_None.I.do.not.use.these.model.explanation.techniques    1974.0
##                                                                     t value
## (Intercept)                                                          86.562
## explainability.model_None.I.do.not.use.these.model.explanation.techniques  -6.336
##                                                                     Pr(>|t|)
## (Intercept)                                                          < 2e-16
## explainability.model_None.I.do.not.use.these.model.explanation.techniques 2.45e-10
##
## (Intercept)                                                         ***
## explainability.model_None.I.do.not.use.these.model.explanation.techniques ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61850 on 10807 degrees of freedom
## Multiple R-squared:  0.003701,	Adjusted R-squared:  0.003609
## F-statistic: 40.14 on 1 and 10807 DF,  p-value: 2.452e-10
```

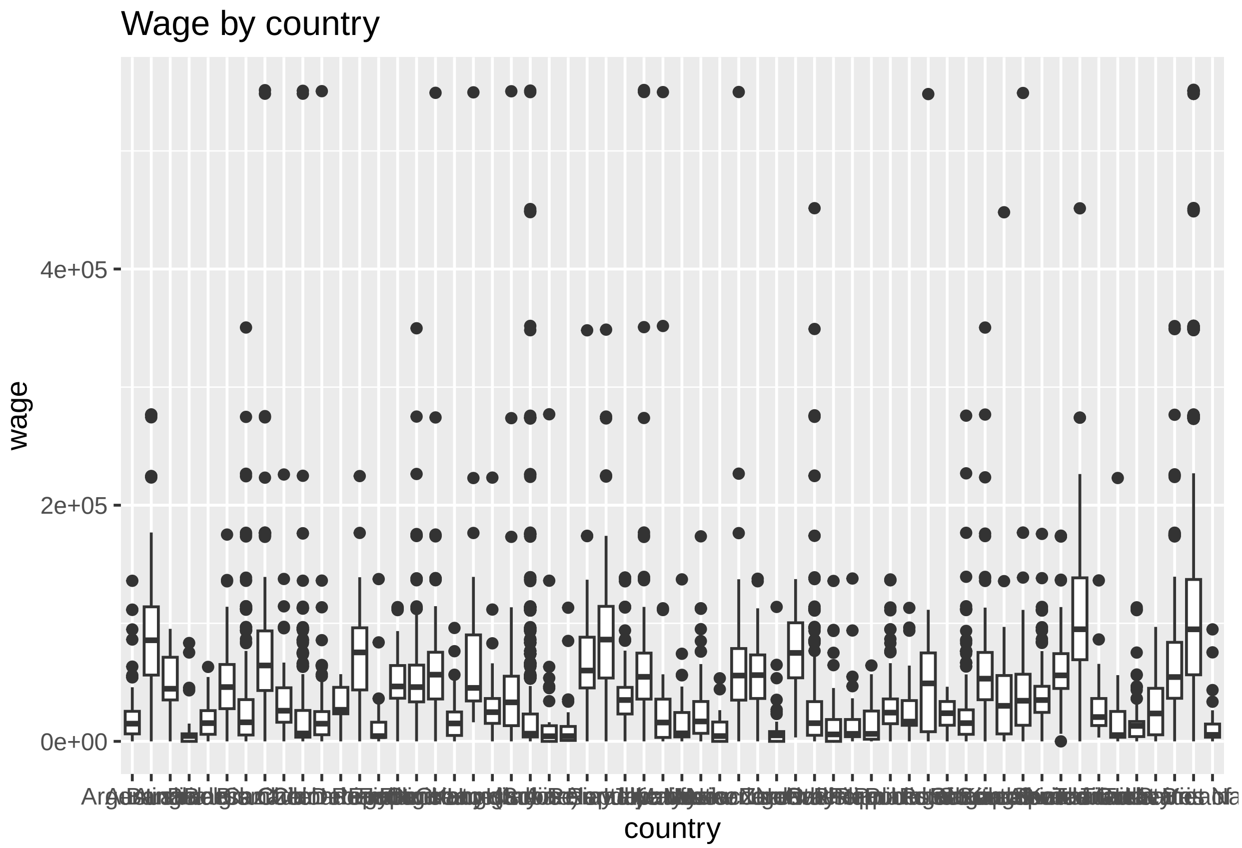
```r
# Group means for categorical features
cat_vars <- c("education", "years_experience", "gender", "country", "job_role", "industry", "age")
for (v in cat_vars) {
  print(ggplot(df, aes_string(x = v, y = "wage")) +
          geom_boxplot() +
          ggtitle(paste("Wage by", v)))
}
```

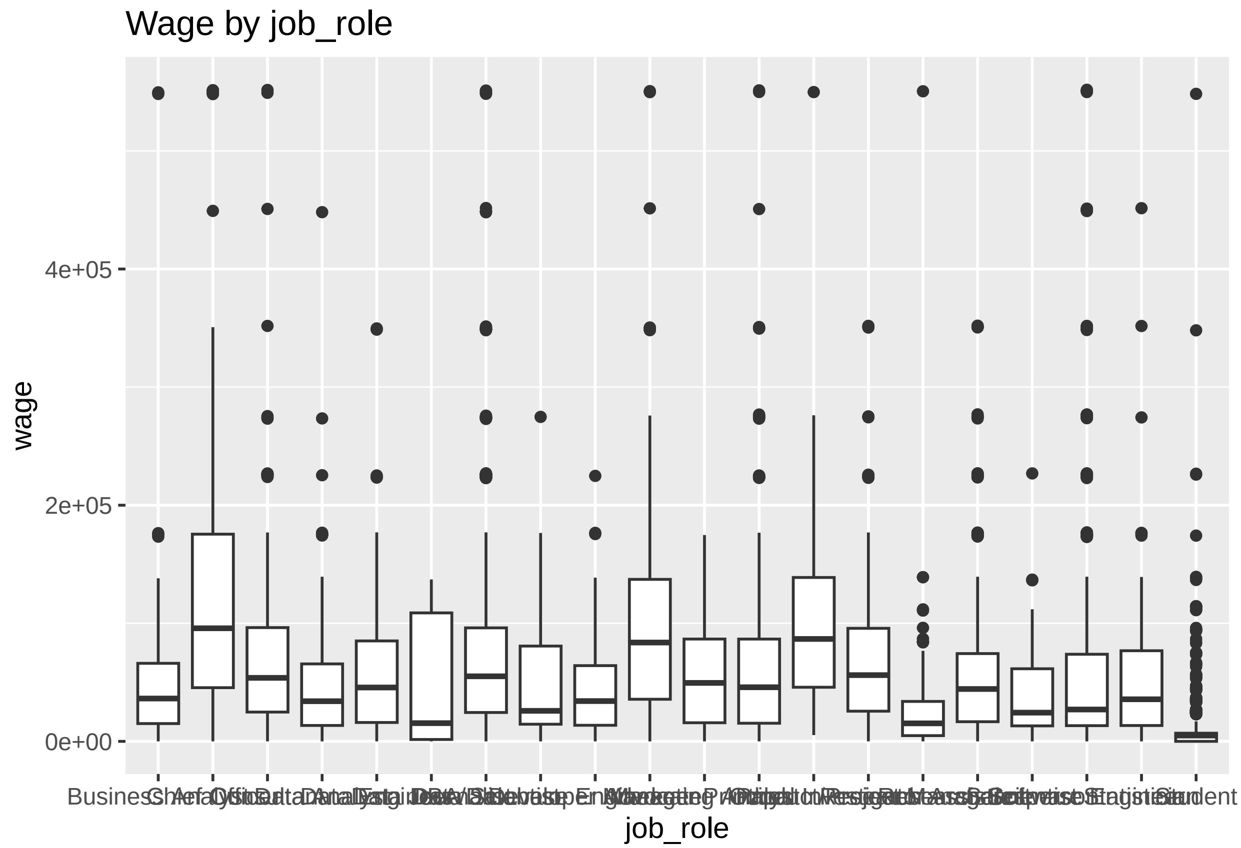


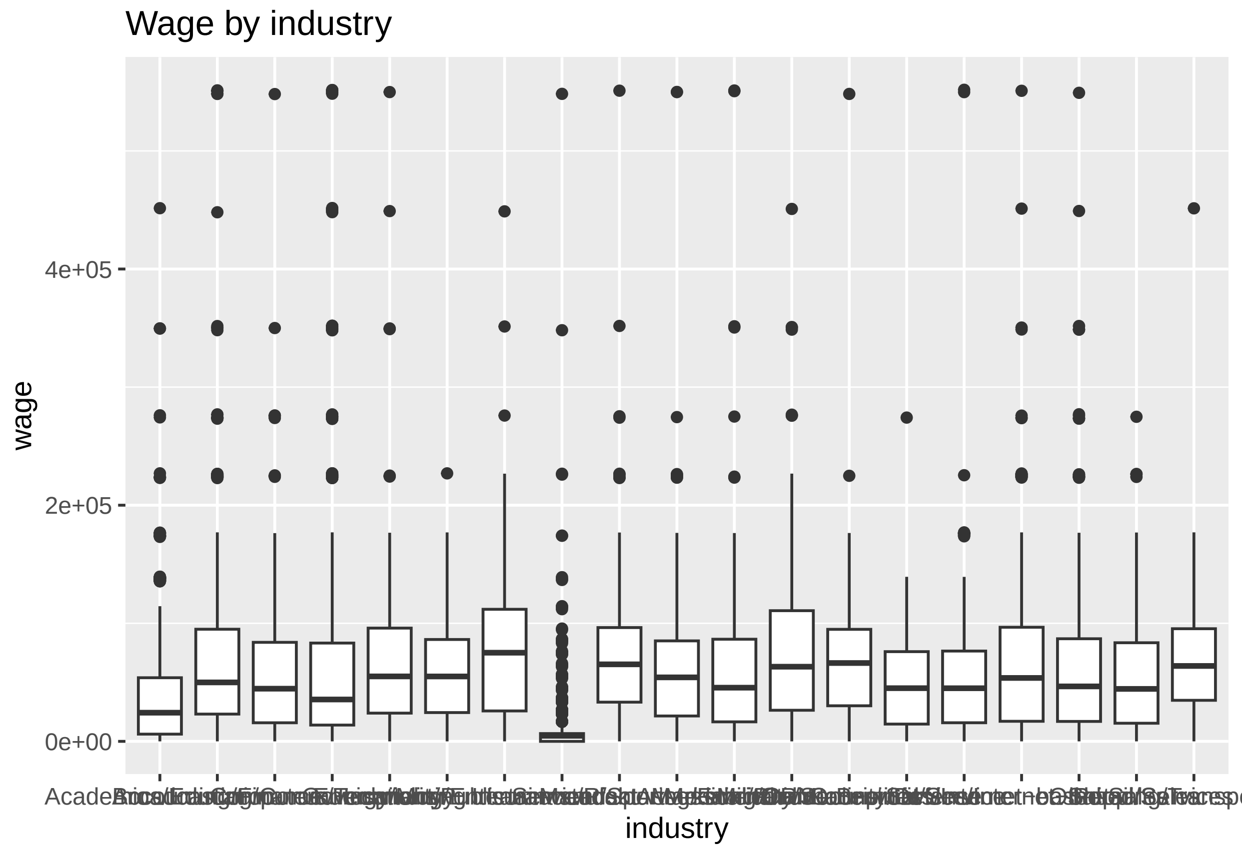
Wage by education
wage
4e+05
2e+05
0e+00
Bachelor’s degree
Doctoral degree
I prefer not to answer
education

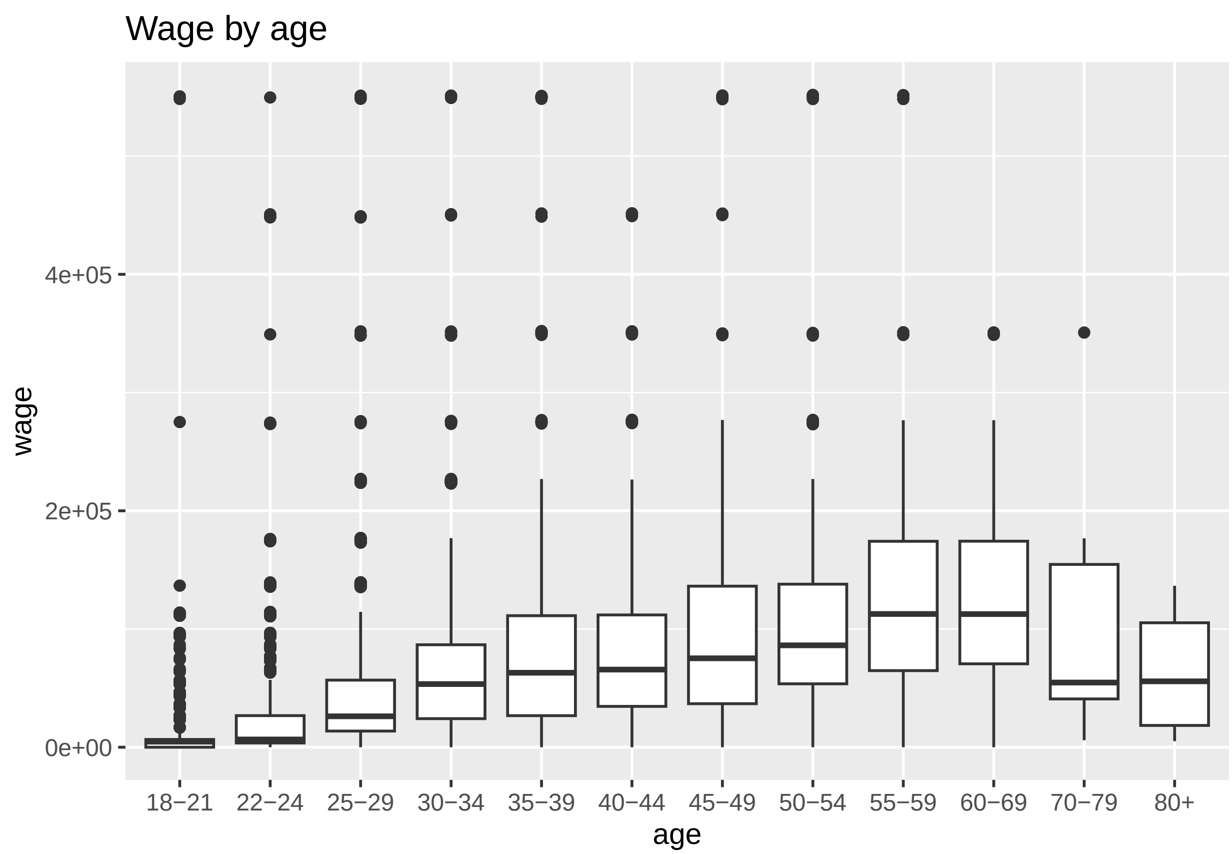

Wage by years_experience
wage
4e+05
2e+05
0e+00
0−1
1−2
11−15
15−20
2−3
20−25
25−30
3−4
30 +
4−5
5−11
years_experience


Wage by gender
wage
4e+05
2e+05
0e+00
Female
Male
Prefer not to say
Prefer to self−describe
gender


Wage by country
wage
4e+05
2e+05
0e+00
country


Wage by job_role
wage
4e+05
2e+05
0e+00
job_role


Wage by industry
wage
4e+05
2e+05
0e+00
industry



5 Data preparation & encoding

Convert categorical variables to factors:

Categorical variables are converted to factors to prepare them for modeling.

```
df$gender <- as.factor(df$gender)
df$education <- as.factor(df$education)
df$country <- as.factor(df$country)
df$age <- as.factor(df$age)
df$years_experience <- as.factor(df$years_experience)
df$job_role <- as.factor(df$job_role)
df$industry <- as.factor(df$industry)
df$ML_atwork <- as.factor(df$ML_atwork)
df$percent_actively.coding <- as.factor(df$percent_actively.coding)
df$For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school.. <- as.factor(df$For
df$How.long.have.you.been.writing.code.to.analyze.data. <- as.factor(df$How.long.have.you.been.writing.
```

6 Feature Selection

We select key predictors and that we think are important for our model to predict the right wage. From the data exploration above we chose the features: age, years_experience, education, gender, country, job_role, industry, ML_atwork, percent_actively.coding, How.long.have.you.been.writing.code.to.analyze.data., For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school..

```r
model_data <- df %>%
  select(wage, age, years_experience, education, gender, country, job_role, industry, ML_atwork, percen
```

Dummy Encoding

We apply dummy encoding to convert categorical features to numeric format for autoML modeling.

```r
dummy_model <- caret::dummyVars(~ ., data = model_data[,-1])
dummy_data <- predict(dummy_model, newdata = model_data[,-1])
model_matrix <- data.frame(wage = model_data$wage, dummy_data)
```

7. AutoML with H2O

We use H2O's AutoML to automatically train and tune multiple models. XGBoost is excluded due to compatibility issues with windows.

```r
h2o.init()
```

```
##  Connection successful!
##
## R is connected to the H2O cluster:
##     H2O cluster uptime:         2 hours 56 minutes
##     H2O cluster timezone:       Europe/Brussels
##     H2O data parsing timezone:  UTC
##     H2O cluster version:        3.44.0.3
##     H2O cluster version age:    1 year, 5 months and 6 days
##     H2O cluster name:           H2O_started_from_R_ianho_bzp644
##     H2O cluster total nodes:    1
##     H2O cluster total memory:   1.51 GB
##     H2O cluster total cores:    8
##     H2O cluster allowed cores:  8
##     H2O cluster healthy:        TRUE
##     H2O Connection ip:          localhost
##     H2O Connection port:        54321
##     H2O Connection proxy:       NA
##     H2O Internal Security:      FALSE
##     R Version:                  R version 4.4.1 (2024-06-14 ucrt)
```

```r
h2o.xgboost.available()
```

```
## [1] "Cannot build a XGboost model - no backend found."
```

```
## [1] FALSE
```

```r
df_h2o <- as.h2o(model_matrix)
```

```
##   |                                                                              |
```

```r
set.seed(12)
splits <- h2o.splitFrame(df_h2o, ratios = 0.8, seed = 1234)
train <- splits[[1]]
valid <- splits[[2]]

dep_var <- "wage"
indep_vars <- setdiff(colnames(df_h2o), dep_var)

automl <- h2o.automl(
  x = indep_vars,
  y = dep_var,
  training_frame = train,
  leaderboard_frame = valid,
  max_models = 13,
  seed = 12,
  sort_metric = "RMSE",
  exclude_algos = c("XGBoost")
)
```

```
##   |                                                                      |
##   H2O connection has been severed. Cannot connect to instance at http://localhost:54321/
## Timeout was reached [localhost]: Connection timeout after 36810 ms
## Error in .h2o.doSafeREST(h2oRestApiVersion = h2oRestApiVersion, urlSuffix = urlSuffix,  :
##   Unexpected CURL error: Timeout was reached [localhost]: Connection timeout after 13837 ms
## [1] "Job request failed Unexpected CURL error: Timeout was reached [localhost]: Connection timeout a
##   |                                                                      |==================
```

```r
lb <- automl@leaderboard
print(lb)
```

```
##                                                    model_id     rmse        mse
## 1    StackedEnsemble_AllModels_1_AutoML_3_20250526_203212 40345.91 1627792817
## 2 StackedEnsemble_BestOfFamily_1_AutoML_3_20250526_203212 40373.98 1630058011
## 3             GBM_grid_1_AutoML_3_20250526_203212_model_2 40789.70 1663799557
## 4                          GBM_4_AutoML_3_20250526_203212 40979.24 1679298083
## 5                          GBM_3_AutoML_3_20250526_203212 41022.45 1682841025
## 6    DeepLearning_grid_1_AutoML_3_20250526_203212_model_1 41168.69 1694861151
##        mae    rmsle mean_residual_deviance
## 1 23030.66 3.026653             1627792817
## 2 22029.96      NaN             1630058011
## 3 22539.05      NaN             1663799557
## 4 22447.41      NaN             1679298083
## 5 22578.69      NaN             1682841025
## 6 23189.42      NaN             1694861151
##
## [15 rows x 6 columns]
```

```r
best_model <- automl@leader
print(best_model)
```

```
## Model Details:
## ==============
```

```
##
## H2ORegressionModel: stackedensemble
## Model ID:  StackedEnsemble_AllModels_1_AutoML_3_20250526_203212
## Model Summary for Stacked Ensemble:
##                                          key           value
## 1                          Stacking strategy cross_validation
## 2       Number of base models (used / total)            9/13
## 3           # GBM base models (used / total)             6/7
## 4  # DeepLearning base models (used / total)             3/3
## 5           # DRF base models (used / total)             0/2
## 6           # GLM base models (used / total)             0/1
## 7                      Metalearner algorithm             GLM
## 8         Metalearner fold assignment scheme          Random
## 9                         Metalearner nfolds               5
## 10                   Metalearner fold_column              NA
## 11        Custom metalearner hyperparameters            None
##
##
## H2ORegressionMetrics: stackedensemble
## ** Reported on training data. **
##
## MSE:  1423449845
## RMSE:  37728.63
## MAE:  20918.41
## RMSLE:  NaN
## Mean Residual Deviance :  1423449845
##
##
##
## H2ORegressionMetrics: stackedensemble
## ** Reported on cross-validation data. **
## ** 5-fold cross-validation on training data (Metrics computed for combined holdout predictions) **
##
## MSE:  1766152367
## RMSE:  42025.62
## MAE:  22128.7
## RMSLE:  NaN
## Mean Residual Deviance :  1766152367
##
##
## Cross-Validation Metrics Summary:
##                                     mean                 sd
## mae                         22156.371000         828.670530
## mean_residual_deviance    1774074000.000000    324389980.000000
## mse                       1774074000.000000    324389980.000000
## null_deviance          6721294600000.000000 503479570000.000000
## r2                              0.543667           0.043387
## residual_deviance      3070473800000.000000 454112120000.000000
## rmse                        41985.950000        3750.704800
## rmsle                                 NA           0.000000
##                               cv_1_valid          cv_2_valid
## mae                         21836.725000        22361.455000
## mean_residual_deviance    1763980200.000000    1585879800.000000
## mse                       1763980200.000000    1585879800.000000
```

```
## null_deviance          7269845000000.000000 6205226000000.000000
## r2                          0.563887             0.554027
## residual_deviance      3164580300000.000000 2767360400000.000000
## rmse                       41999.766000         39823.105000
## rmsle                            NA                   NA
##                          cv_3_valid           cv_4_valid
## mae                        21064.920000         22174.550000
## mean_residual_deviance    1443141500.000000    1778066800.000000
## mse                       1443141500.000000    1778066800.000000
## null_deviance          6314774000000.000000 6583105400000.000000
## r2                          0.597128             0.519312
## residual_deviance      2534156500000.000000 3161402600000.000000
## rmse                       37988.703000         42167.130000
## rmsle                            NA                   NA
##                          cv_5_valid
## mae                        23344.205000
## mean_residual_deviance    2299302000.000000
## mse                       2299302000.000000
## null_deviance          7233522000000.000000
## r2                          0.483983
## residual_deviance      3724869000000.000000
## rmse                       47951.035000
## rmsle                            NA
```

8. Random Forest Model Development

We build a traditional Random Forest for comparison.

We split the data into training and testing sets to evaluate model generalization.

We chose Random Forest as a robust, interpretable, and non-parametric model that performs well on high-dimensional, categorical datasets like ours which has over 70 categorical features. It served as a reliable benchmark against AutoML's more complex ensembles while keeping control over the modeling process.

```r
set.seed(123)

train_index <- createDataPartition(model_matrix$wage, p = 0.7, list = FALSE)
train_data <- model_matrix[train_index, ]
test_data  <- model_matrix[-train_index, ]


rf_model <- randomForest(wage ~ ., data = train_data, importance = TRUE)
print(rf_model)
```

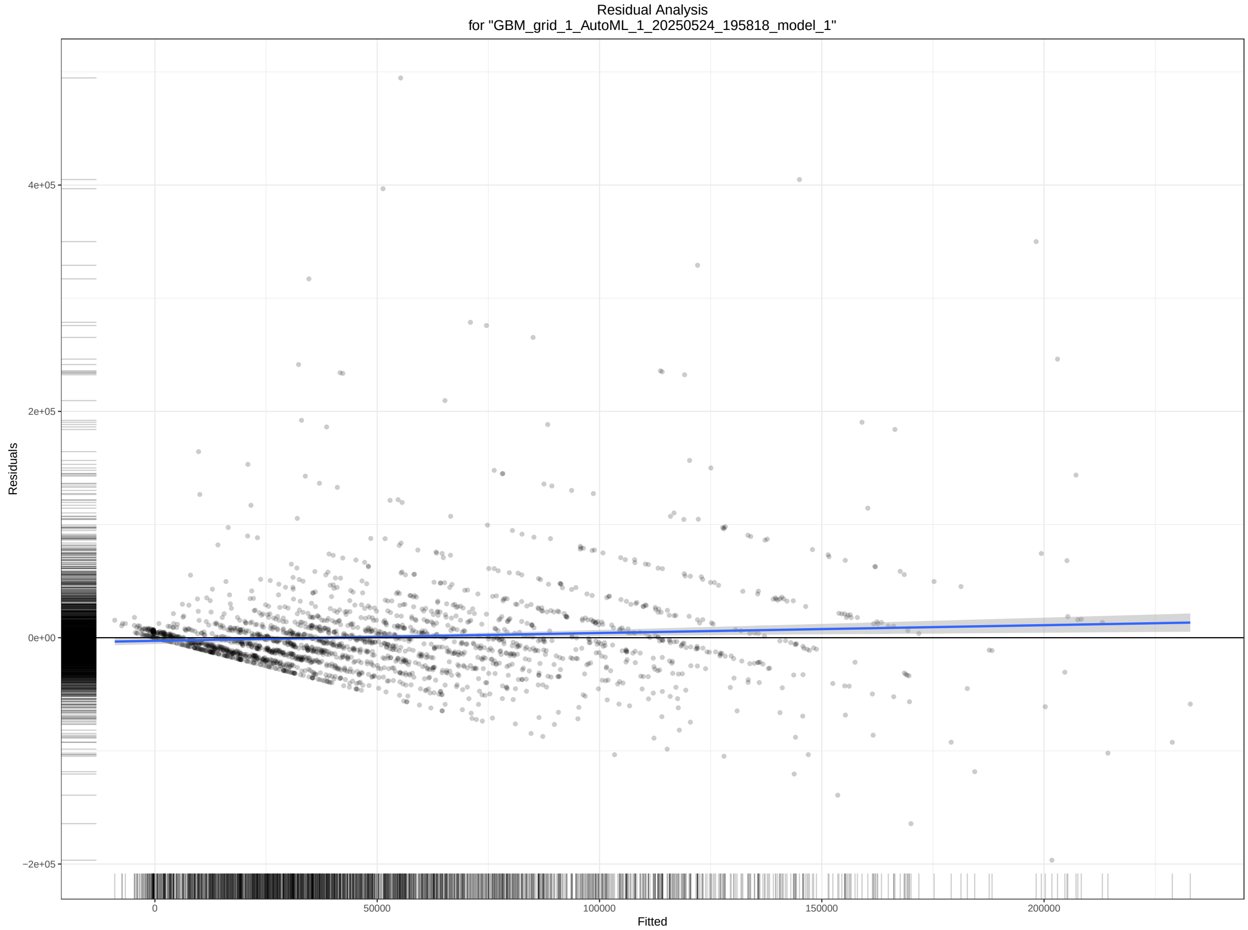
```
##
## Call:
##  randomForest(formula = wage ~ ., data = train_data, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 54
##
##           Mean of squared residuals: 1865149410
##                     % Var explained: 51.04
```

9 AutoML Explainability

We use H2O's built-in tools to explore variable importance and local explanations.

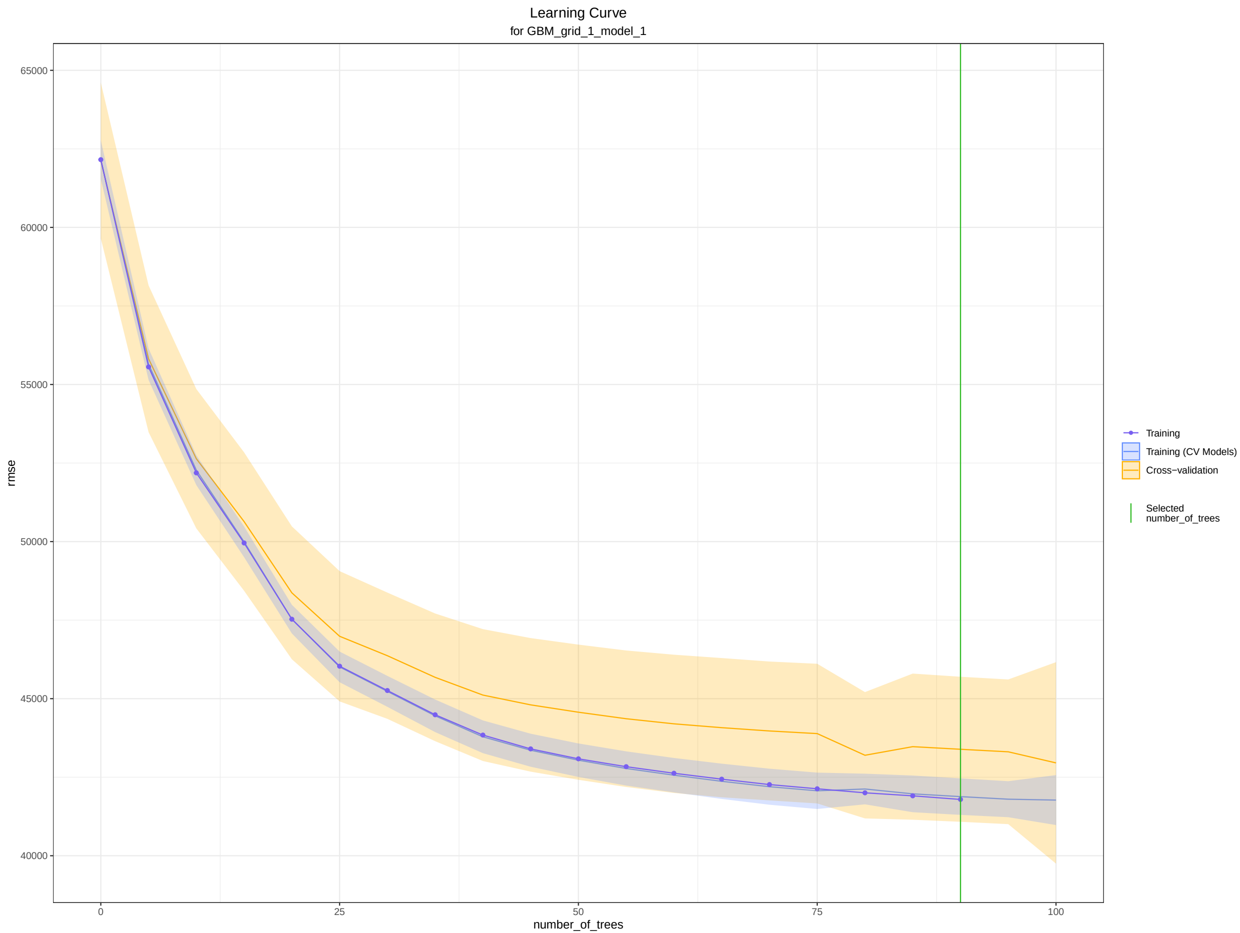
```
StackedEnsemble_rsme <- h2o.get_best_model(automl, algorithm = "StackedEnsemble", criterion = "rmse")
GBM_rsme <- h2o.get_best_model(automl, algorithm = "GBM", criterion = "rmse")

### We do this but outside of knitting: exp_GBM <- h2o.explain(GBM_rsme, valid)
### save(exp_GBM, file = "exp_GBM.RData")

load("exp_GBM.RData")
print(exp_GBM)
```

```
##
##
## Residual Analysis
## =================
##
## > Residual Analysis plots the fitted values vs residuals on a test dataset. Ideally, residuals shoul
```

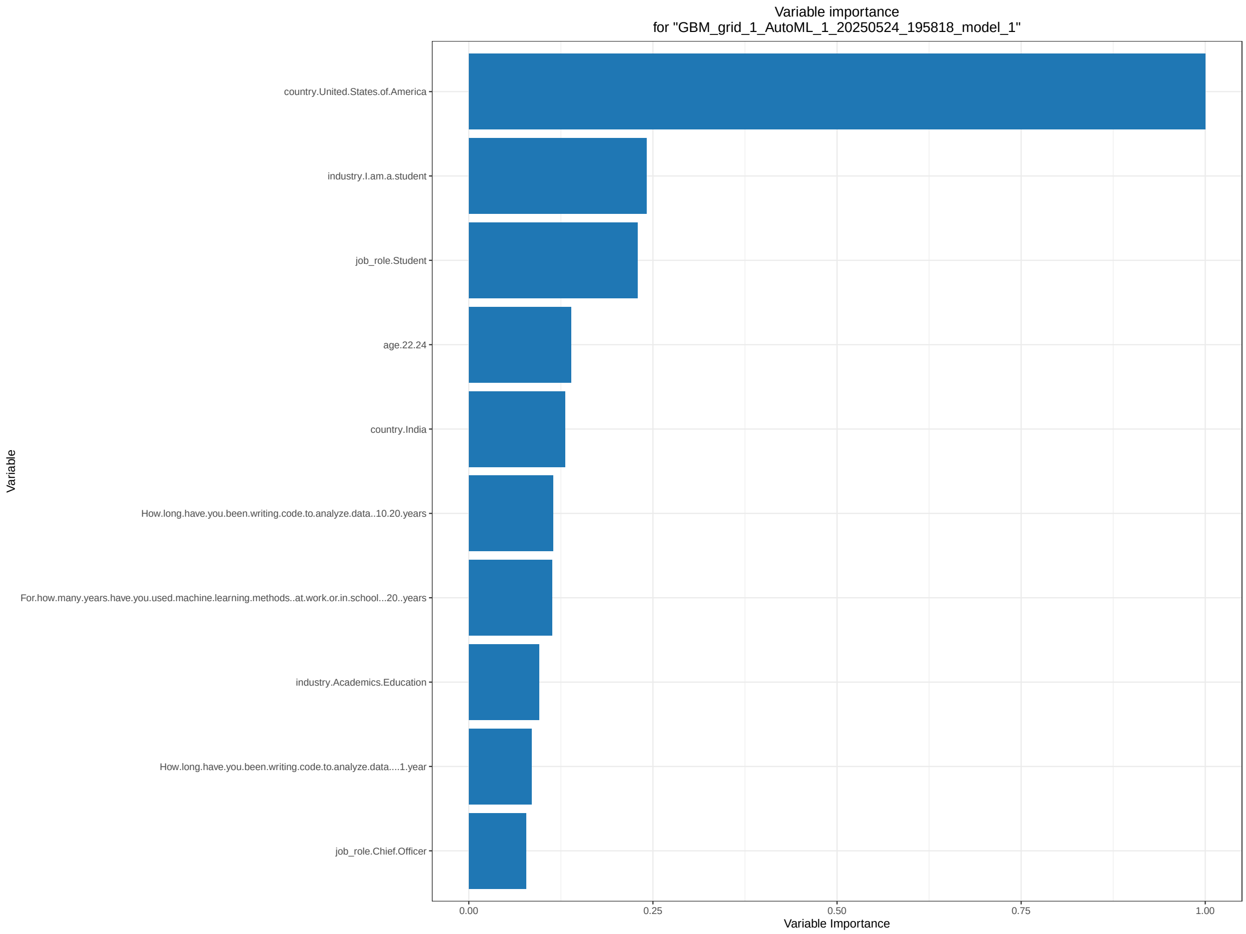



```
##
##
```

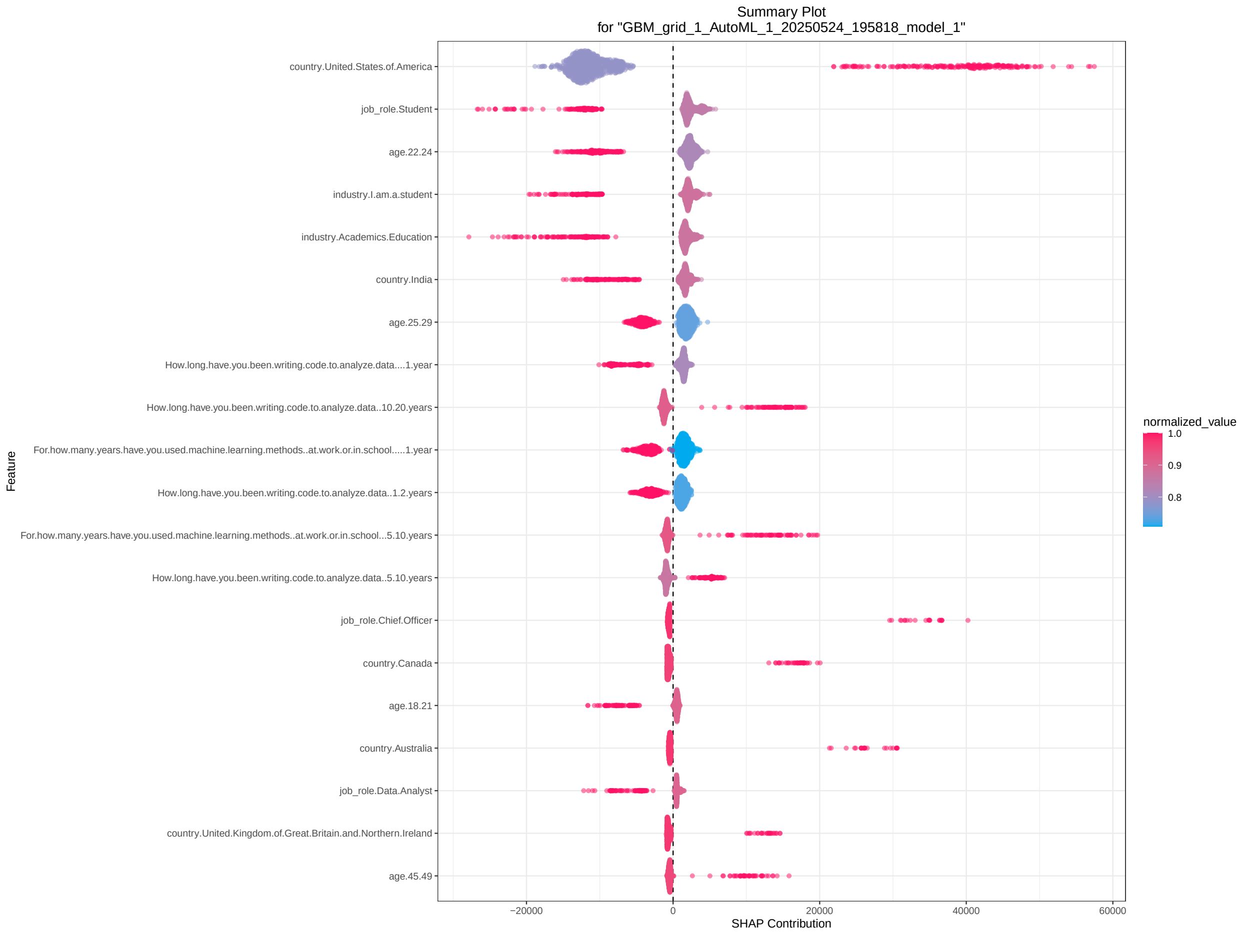
```
## Learning Curve Plot
## ===================
##
## > Learning curve plot shows the loss function/metric dependent on number of iterations or trees for t
```



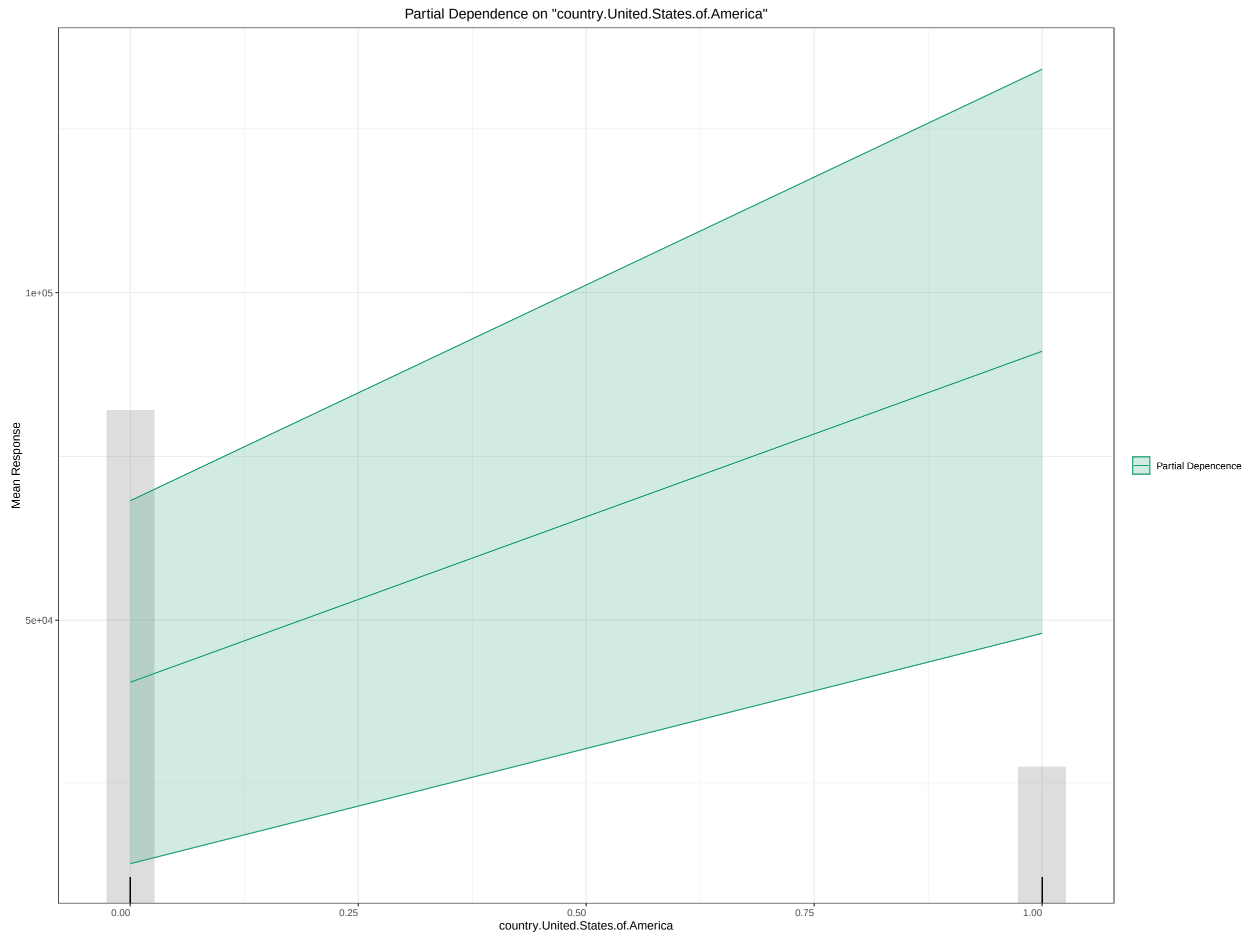

```
##
##
## Variable Importance
## ===================
##
## > The variable importance plot shows the relative importance of the most important variables in the m
```

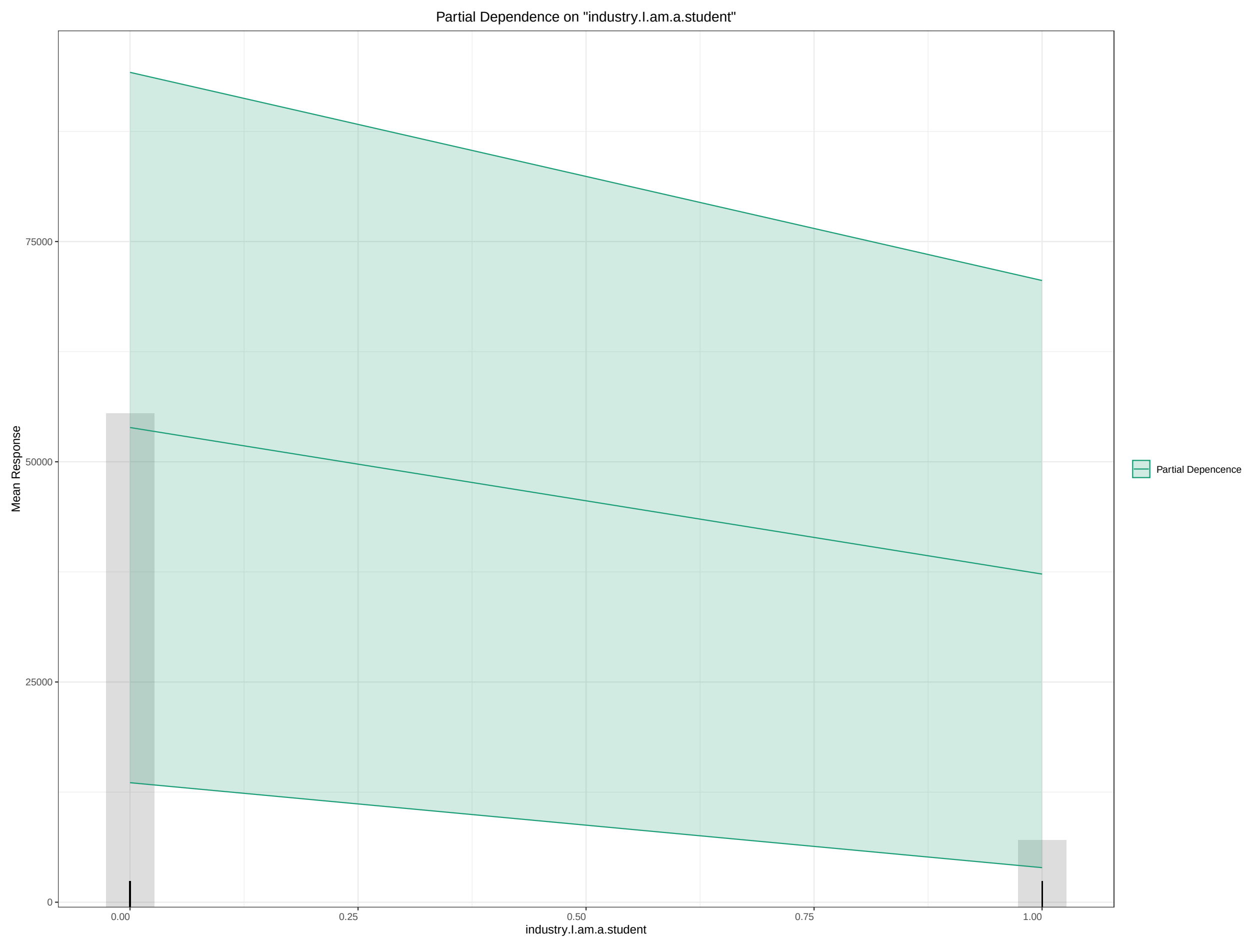


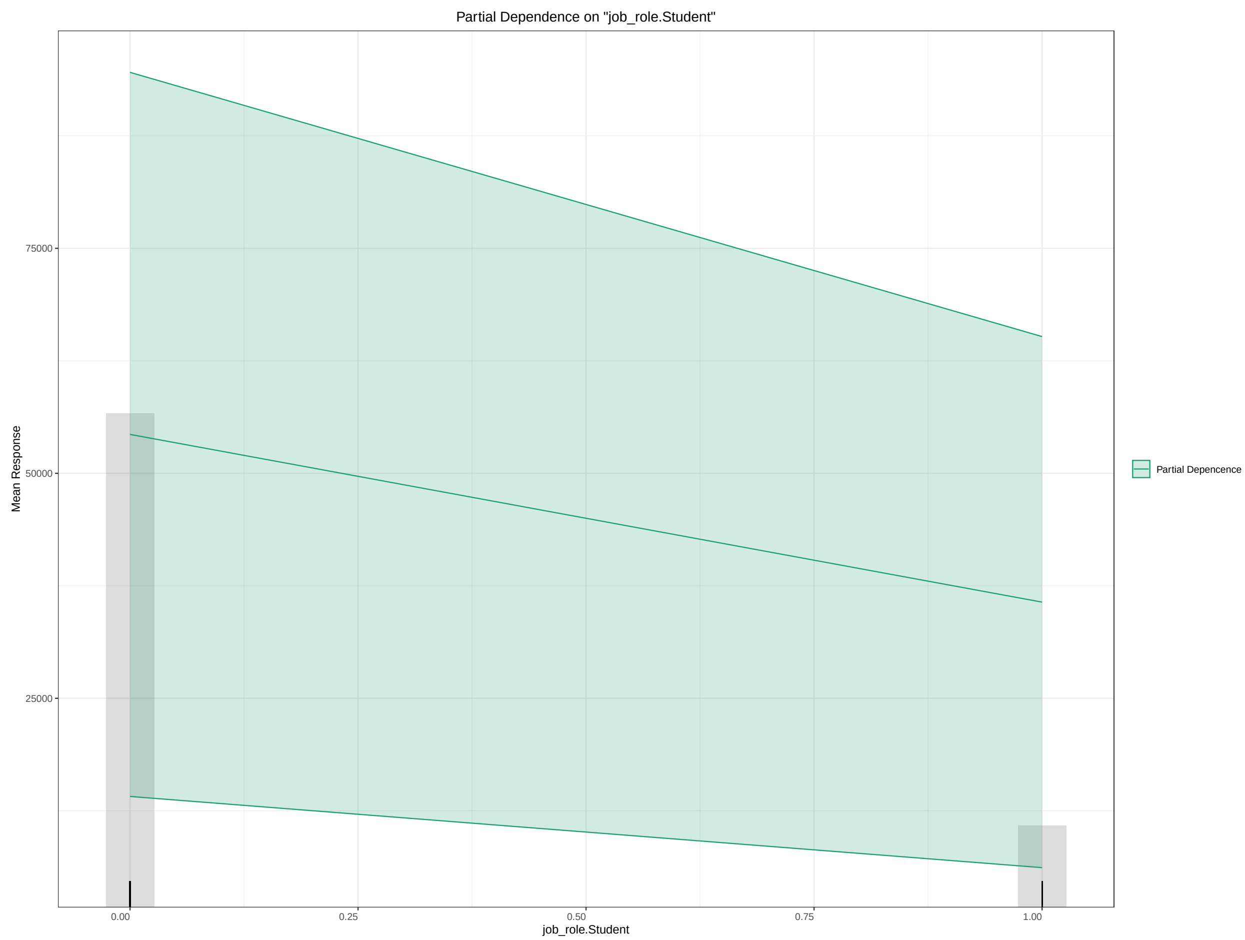

```
##
##
## SHAP Summary
## ============
##
## > SHAP summary plot shows the contribution of the features for each instance (row of data). The sum o
```

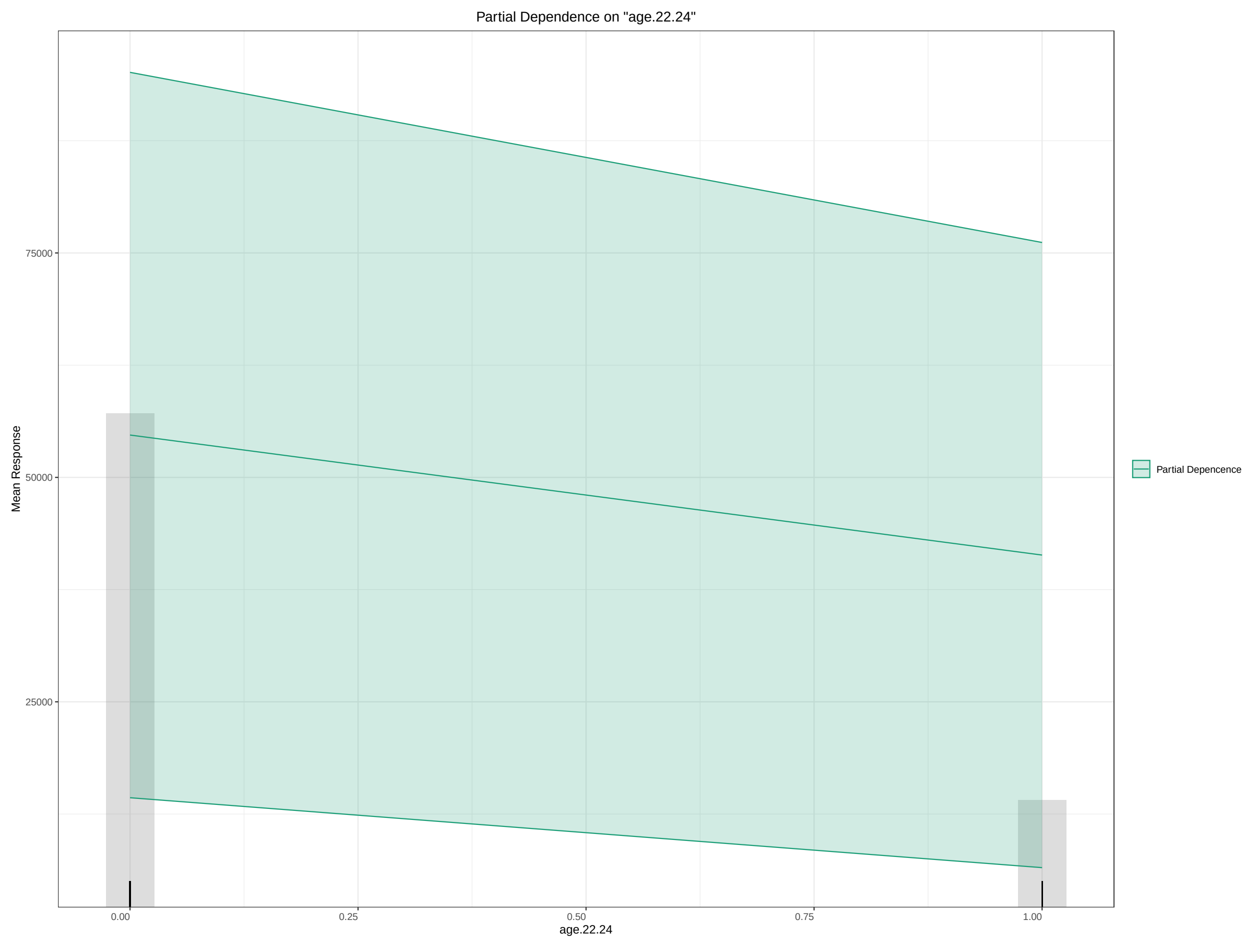



```
##
##
## Partial Dependence Plots
## ========================
##
## > Partial dependence plot (PDP) gives a graphical depiction of the marginal effect of a variable on t
```

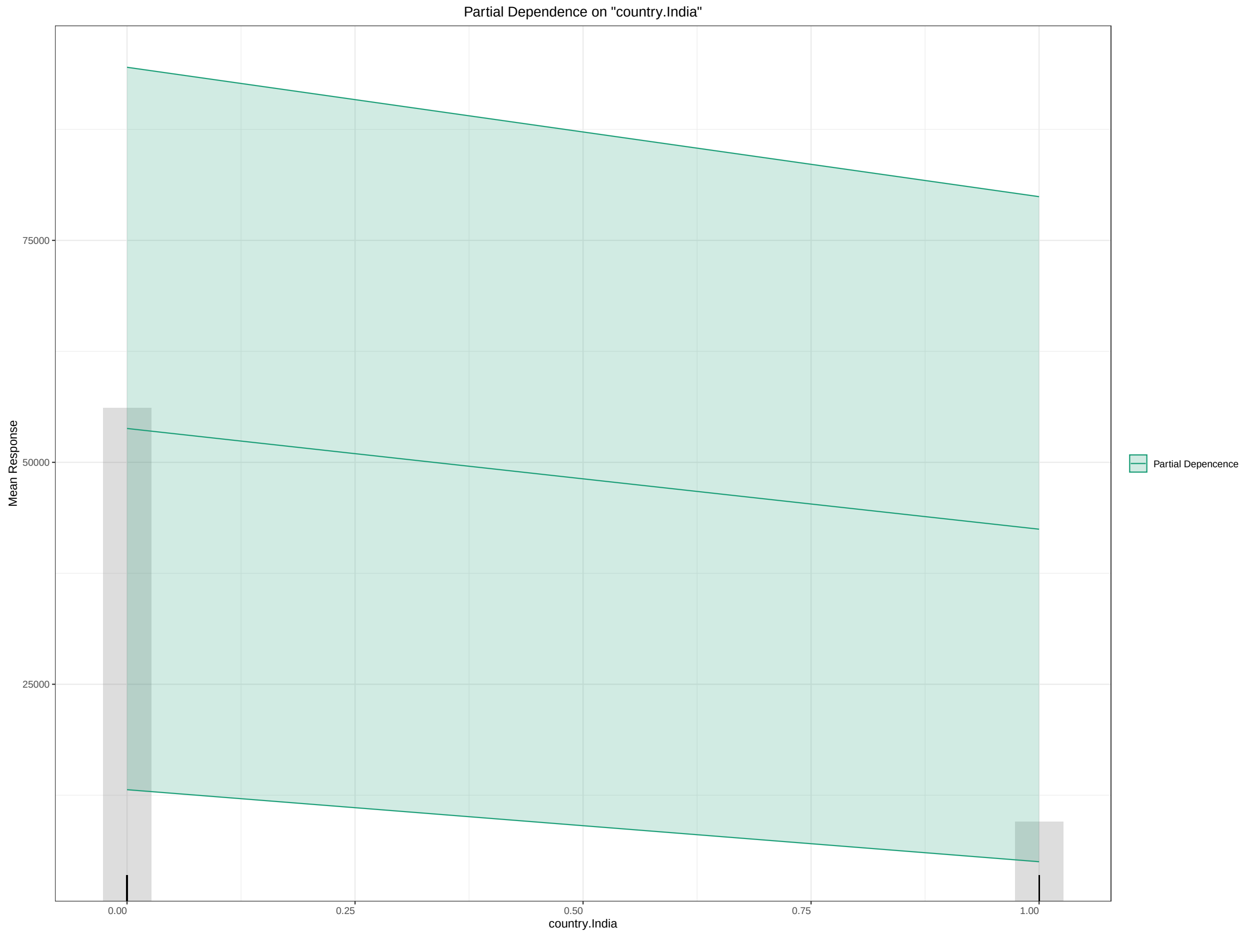

Partial Dependence on "country.United.States.of.America"
Mean Response
1e+05
5e+04
0.00
0.25
0.50
0.75
1.00
country.United.States.of.America
Partial Depencence


Partial Dependence on "industry.I.am.a.student"
Mean Response
75000
50000
25000
0
0.00
0.25
0.50
0.75
1.00
industry.I.am.a.student
Partial Depencence

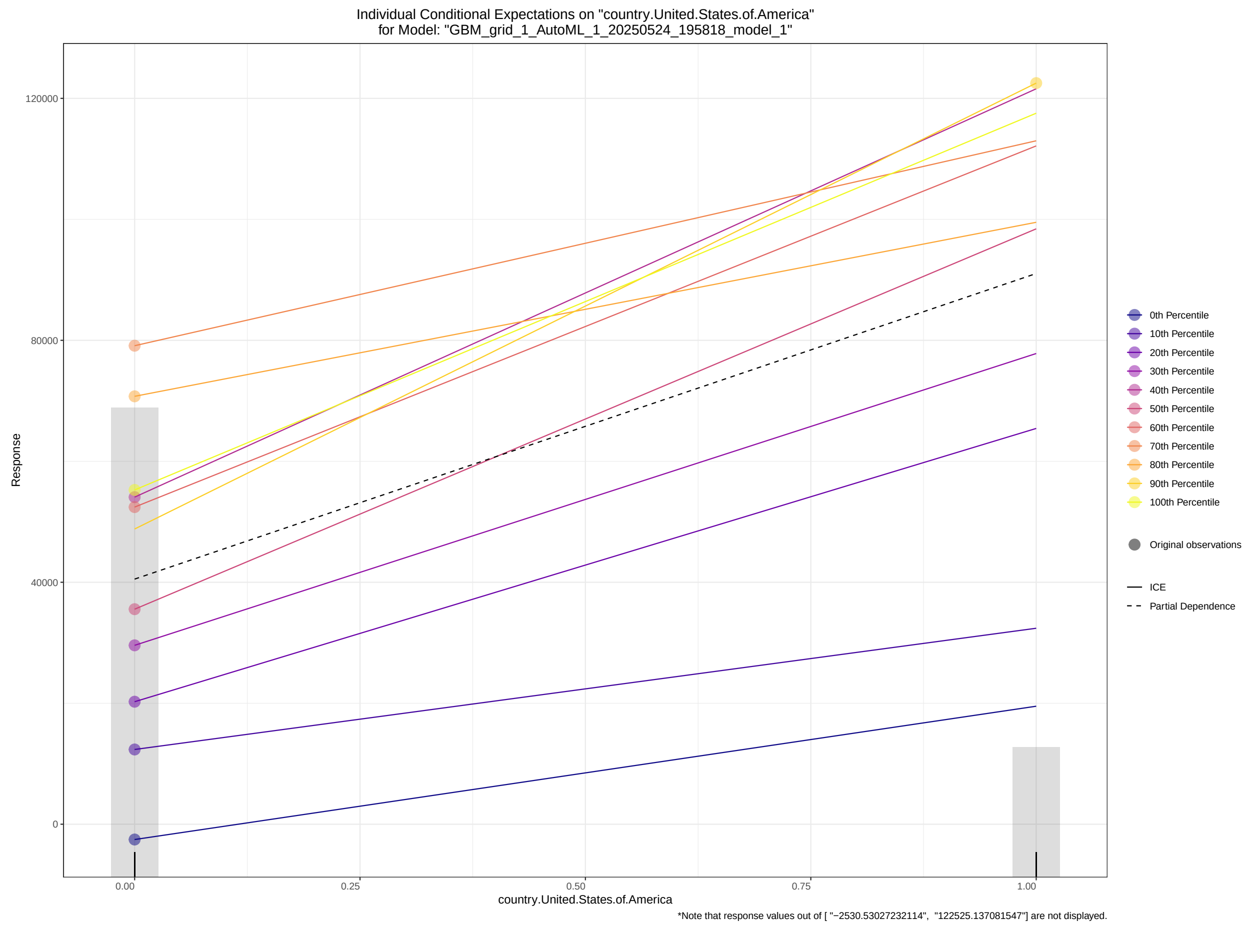

Partial Dependence on "job_role.Student"
Mean Response
75000
50000
25000
0.00
0.25
0.50
0.75
1.00
job_role.Student
Partial Depencence

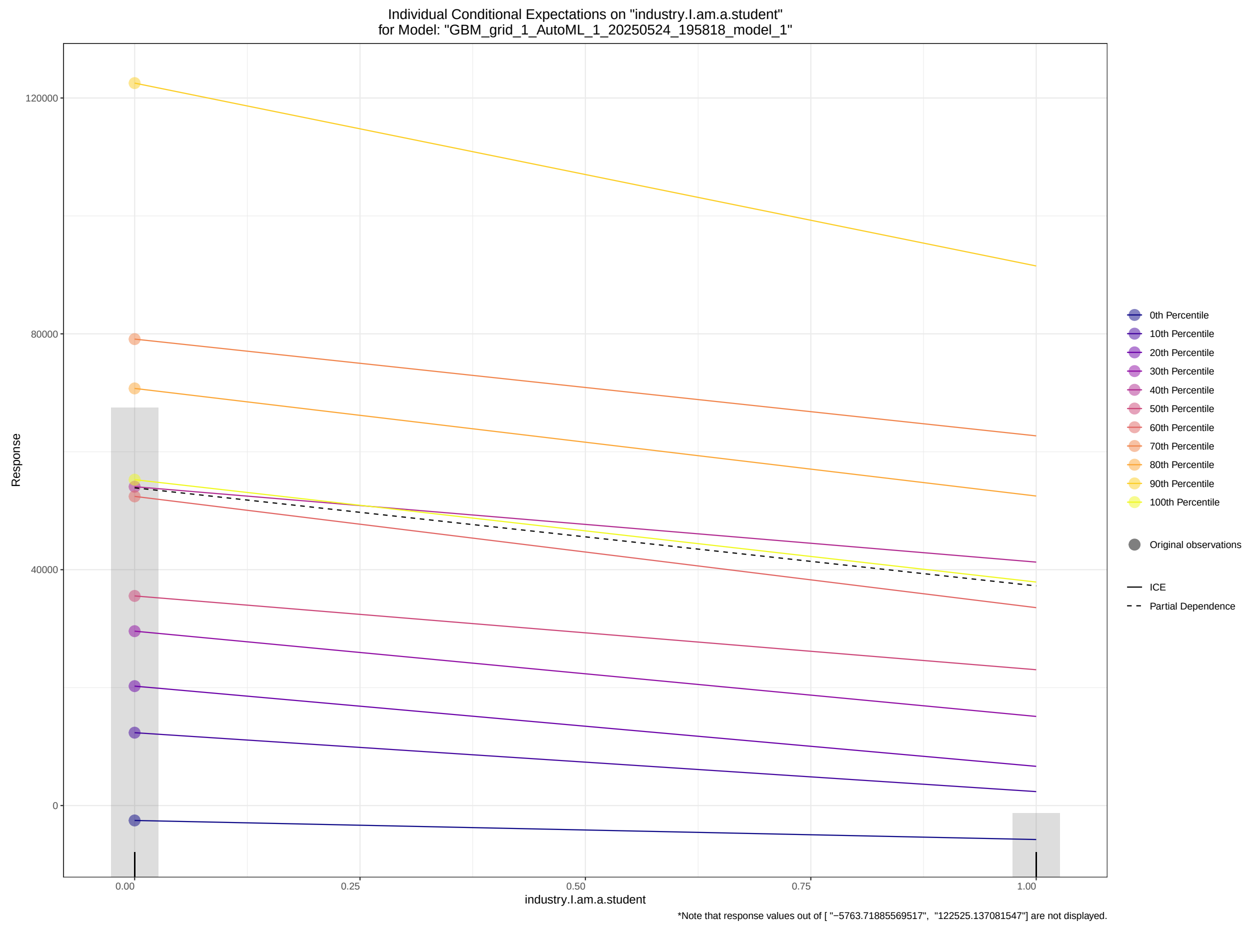


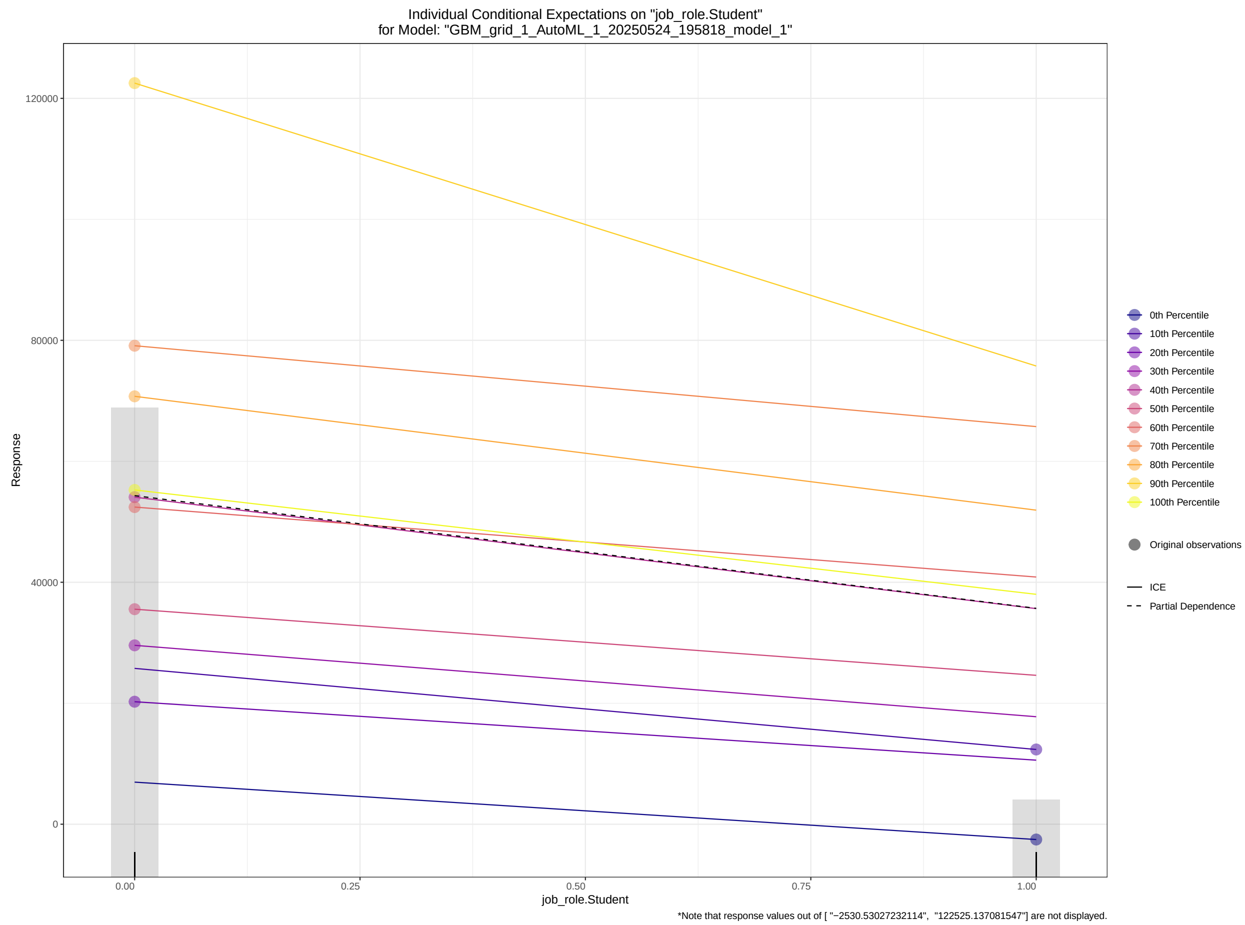
Partial Dependence on "age.22.24"
Mean Response
75000
50000
25000
0.00
0.25
0.50
0.75
1.00
age.22.24
Partial Depencence



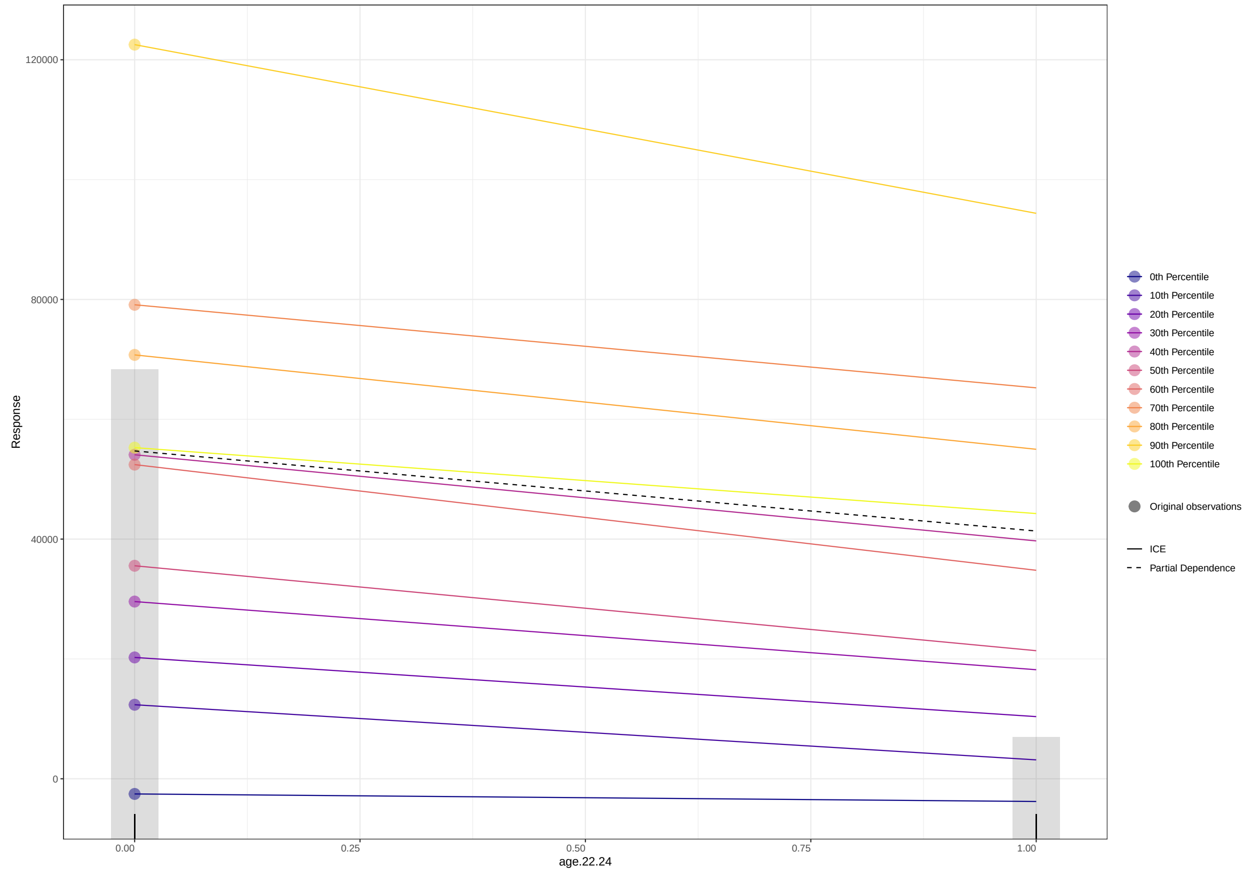

```
##
##
## Individual Conditional Expectations
## ===================================
##
## > An Individual Conditional Expectation (ICE) plot gives a graphical depiction of the marginal effec
```


Individual Conditional Expectations on "country.United.States.of.America"
for Model: "GBM_grid_1_AutoML_1_20250524_195818_model_1"
Response
120000
80000
40000
0
0.00
0.25
0.50
0.75
1.00
country.United.States.of.America
*Note that response values out of ["−2530.53027232114", "122525.137081547"] are not displayed.
0th Percentile
10th Percentile
20th Percentile
30th Percentile
40th Percentile
50th Percentile
60th Percentile
70th Percentile
80th Percentile
90th Percentile
100th Percentile
Original observations
ICE
Partial Dependence

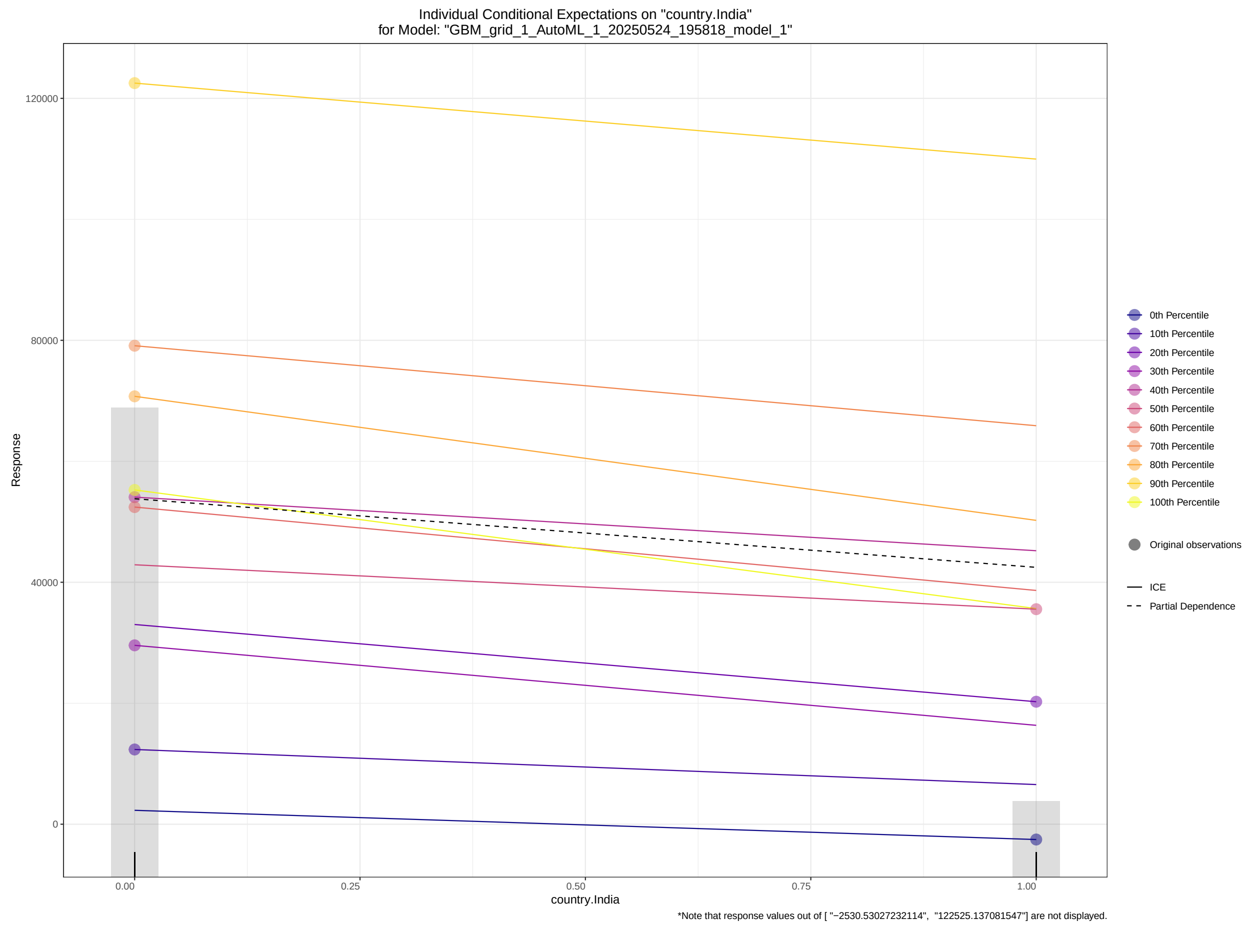

Individual Conditional Expectations on "industry.I.am.a.student"
for Model: "GBM_grid_1_AutoML_1_20250524_195818_model_1"
Response
120000
80000
40000
0
0.00
0.25
0.50
0.75
1.00
industry.I.am.a.student
*Note that response values out of ["−5763.71885569517", "122525.137081547"] are not displayed.
0th Percentile
10th Percentile
20th Percentile
30th Percentile
40th Percentile
50th Percentile
60th Percentile
70th Percentile
80th Percentile
90th Percentile
100th Percentile
Original observations
ICE
Partial Dependence


Individual Conditional Expectations on "job_role.Student"
for Model: "GBM_grid_1_AutoML_1_20250524_195818_model_1"
Response
120000
80000
40000
0
0.00
0.25
0.50
0.75
1.00
job_role.Student
0th Percentile
10th Percentile
20th Percentile
30th Percentile
40th Percentile
50th Percentile
60th Percentile
70th Percentile
80th Percentile
90th Percentile
100th Percentile
Original observations
ICE
Partial Dependence
*Note that response values out of ["−2530.53027232114", "122525.137081547"] are not displayed.

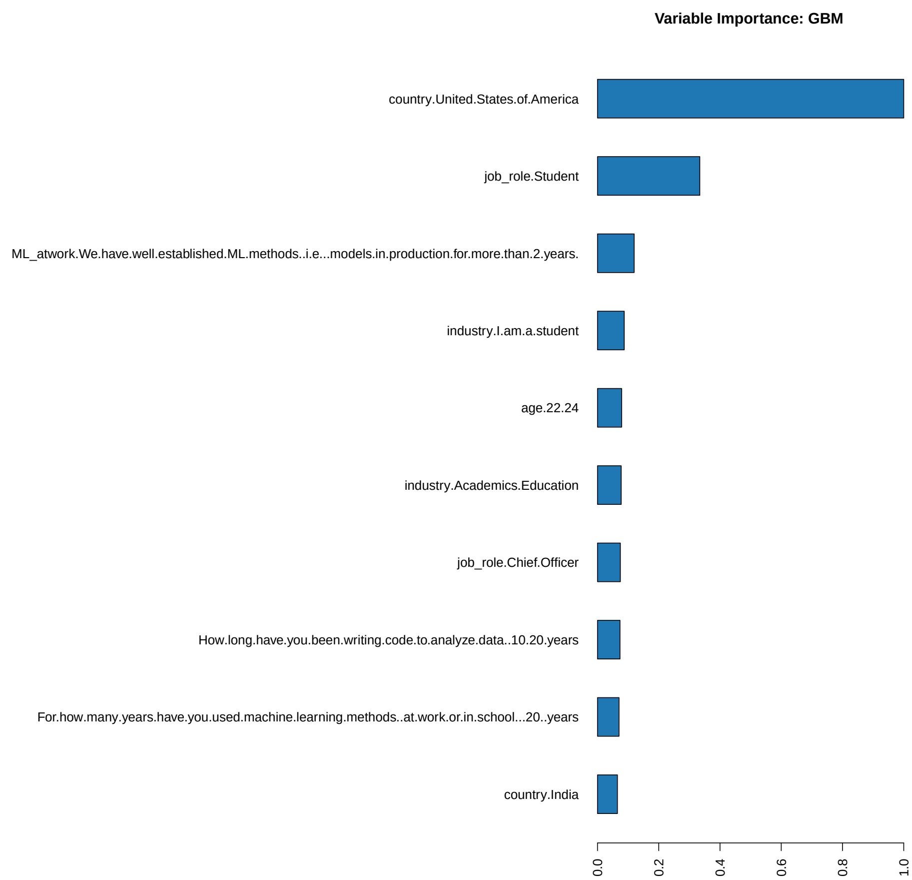
Individual Conditional Expectations on "age.22.24"
for Model: "GBM_grid_1_AutoML_1_20250524_195818_model_1"



*Note that response values out of ["−3789.63874705259", "122525.137081547"] are not displayed.




```r
h2o.varimp_plot(GBM_rsme)
```



10 Random Forest Explainability

We visualize feature importance and partial dependence of wages on experience.

```r
# For Random Forest
importance_values <- importance(rf_model)
print(importance_values)
```

```
##
## age.18.21
## age.22.24
## age.25.29
## age.30.34
## age.35.39
## age.40.44
## age.45.49
## age.50.54
## age.55.59
## age.60.69
## age.70.79
## age.80.
## years_experience.0.1
```

```
## years_experience.1.2
## years_experience.11.15
## years_experience.15.20
## years_experience.2.3
## years_experience.20.25
## years_experience.25.30
## years_experience.3.4
## years_experience.30..
## years_experience.4.5
## years_experience.5.11
## education.Bachelor.s.degree
## education.Doctoral.degree
## education.I.prefer.not.to.answer
## education.Master.s.degree
## education.Professional.degree
## education.Some.college.university.study.without.earning.a.bachelor.s.degree
## gender.Female
## gender.Male
## gender.Prefer.not.to.say
## gender.Prefer.to.self.describe
## country.Argentina
## country.Australia
## country.Austria
## country.Bangladesh
## country.Belarus
## country.Belgium
## country.Brazil
## country.Canada
## country.Chile
## country.China
## country.Colombia
## country.Czech.Republic
## country.Denmark
## country.Egypt
## country.Finland
## country.France
## country.Germany
## country.Greece
## country.Hong.Kong..S.A.R..
## country.Hungary
## country.I.do.not.wish.to.disclose.my.location
## country.India
## country.Indonesia
## country.Iran..Islamic.Republic.of...
## country.Ireland
## country.Israel
## country.Italy
## country.Japan
## country.Kenya
## country.Malaysia
## country.Mexico
## country.Morocco
## country.Netherlands
## country.New.Zealand
```

```
## country.Nigeria
## country.Norway
## country.Other
## country.Pakistan
## country.Peru
## country.Philippines
## country.Poland
## country.Portugal
## country.Republic.of.Korea
## country.Romania
## country.Russia
## country.Singapore
## country.South.Africa
## country.South.Korea
## country.Spain
## country.Sweden
## country.Switzerland
## country.Thailand
## country.Tunisia
## country.Turkey
## country.Ukraine
## country.United.Kingdom.of.Great.Britain.and.Northern.Ireland
## country.United.States.of.America
## country.Viet.Nam
## job_role.Business.Analyst
## job_role.Chief.Officer
## job_role.Consultant
## job_role.Data.Analyst
## job_role.Data.Engineer
## job_role.Data.Journalist
## job_role.Data.Scientist
## job_role.DBA.Database.Engineer
## job_role.Developer.Advocate
## job_role.Manager
## job_role.Marketing.Analyst
## job_role.Other
## job_role.Principal.Investigator
## job_role.Product.Project.Manager
## job_role.Research.Assistant
## job_role.Research.Scientist
## job_role.Salesperson
## job_role.Software.Engineer
## job_role.Statistician
## job_role.Student
## industry.Academics.Education
## industry.Accounting.Finance
## industry.Broadcasting.Communications
## industry.Computers.Technology
## industry.Energy.Mining
## industry.Government.Public.Service
## industry.Hospitality.Entertainment.Sports
## industry.I.am.a.student
## industry.Insurance.Risk.Assessment
## industry.Manufacturing.Fabrication
```

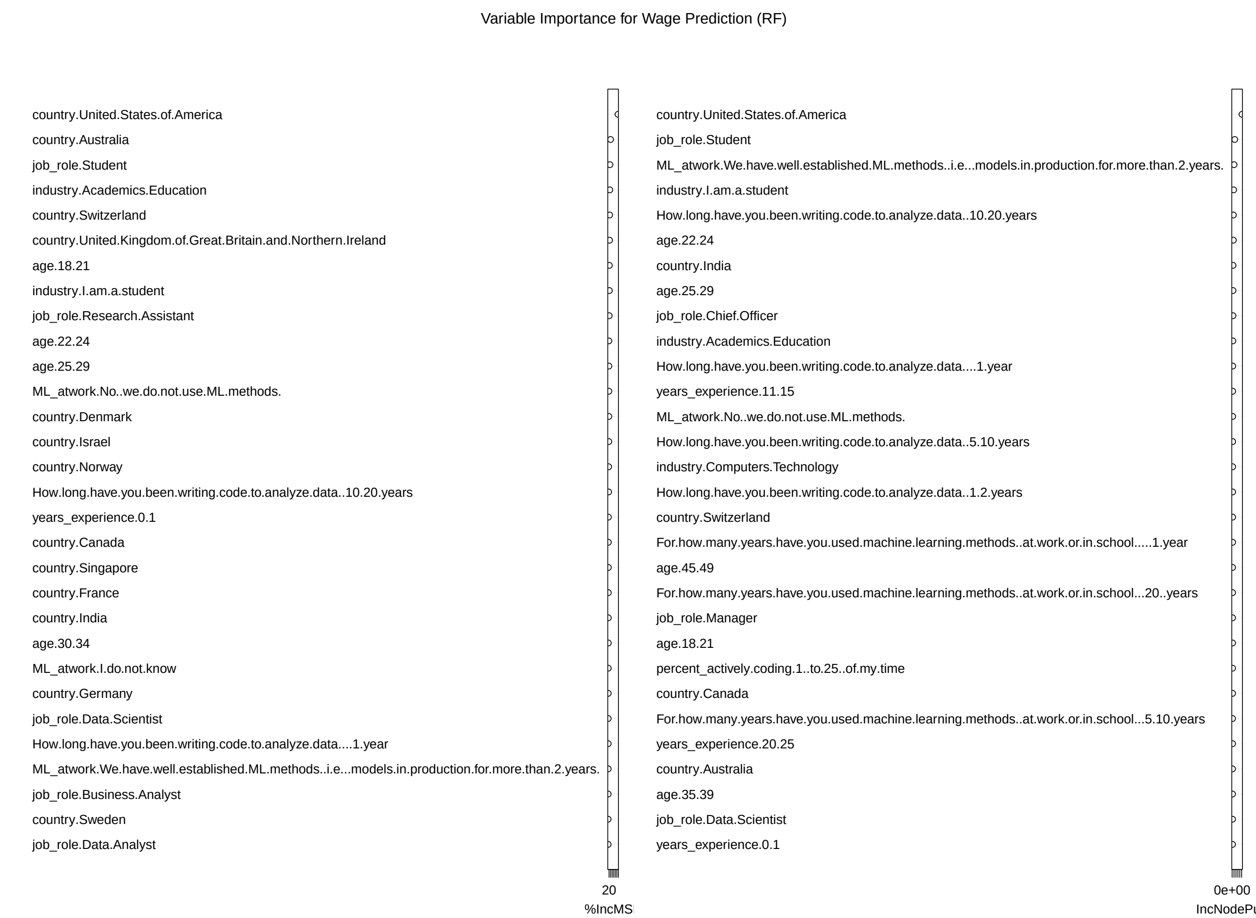
```
## industry.Marketing.CRM
## industry.Medical.Pharmaceutical
## industry.Military.Security.Defense
## industry.Non.profit.Service
## industry.Online.Business.Internet.based.Sales
## industry.Online.Service.Internet.based.Services
## industry.Other
## industry.Retail.Sales
## industry.Shipping.Transportation
## ML_atwork.I.do.not.know
## ML_atwork.No..we.do.not.use.ML.methods.
## ML_atwork.We.are.exploring.ML.methods..and.may.one.day.put.a.model.into.production.
## ML_atwork.We.have.well.established.ML.methods..i.e...models.in.production.for.more.than.2.years.
## ML_atwork.We.recently.started.using.ML.methods..i.e...models.in.production.for.less.than.2.years.
## ML_atwork.We.use.ML.methods.for.generating.insights..but.do.not.put.working.models.into.production.
## percent_actively.coding.0..of.my.time
## percent_actively.coding.1..to.25..of.my.time
## percent_actively.coding.100..of.my.time
## percent_actively.coding.25..to.49..of.my.time
## percent_actively.coding.50..to.74..of.my.time
## percent_actively.coding.75..to.99..of.my.time
## How.long.have.you.been.writing.code.to.analyze.data....1.year
## How.long.have.you.been.writing.code.to.analyze.data..1.2.years
## How.long.have.you.been.writing.code.to.analyze.data..10.20.years
## How.long.have.you.been.writing.code.to.analyze.data..20.30.years
## How.long.have.you.been.writing.code.to.analyze.data..3.5.years
## How.long.have.you.been.writing.code.to.analyze.data..30.40.years
## How.long.have.you.been.writing.code.to.analyze.data..40..years
## How.long.have.you.been.writing.code.to.analyze.data..5.10.years
## How.long.have.you.been.writing.code.to.analyze.data..I.have.never.written.code.and.I.do.not.want.to.l
## How.long.have.you.been.writing.code.to.analyze.data..I.have.never.written.code.but.I.want.to.learn
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school.....1.year
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...1.2.years
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...10.15.years
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...2.3.years
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...20..years
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...3.4.years
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...4.5.years
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...5.10.years
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...I.have.never.studie
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...I.have.never.studie
##
## age.18.21
## age.22.24
## age.25.29
## age.30.34
## age.35.39
## age.40.44
## age.45.49
## age.50.54
## age.55.59
## age.60.69
## age.70.79
## age.80.
```

```
## years_experience.0.1
## years_experience.1.2
## years_experience.11.15
## years_experience.15.20
## years_experience.2.3
## years_experience.20.25
## years_experience.25.30
## years_experience.3.4
## years_experience.30..
## years_experience.4.5
## years_experience.5.11
## education.Bachelor.s.degree
## education.Doctoral.degree
## education.I.prefer.not.to.answer
## education.Master.s.degree
## education.Professional.degree
## education.Some.college.university.study.without.earning.a.bachelor.s.degree
## gender.Female
## gender.Male
## gender.Prefer.not.to.say
## gender.Prefer.to.self.describe
## country.Argentina
## country.Australia
## country.Austria
## country.Bangladesh
## country.Belarus
## country.Belgium
## country.Brazil
## country.Canada
## country.Chile
## country.China
## country.Colombia
## country.Czech.Republic
## country.Denmark
## country.Egypt
## country.Finland
## country.France
## country.Germany
## country.Greece
## country.Hong.Kong..S.A.R..
## country.Hungary
## country.I.do.not.wish.to.disclose.my.location
## country.India
## country.Indonesia
## country.Iran..Islamic.Republic.of...
## country.Ireland
## country.Israel
## country.Italy
## country.Japan
## country.Kenya
## country.Malaysia
## country.Mexico
## country.Morocco
## country.Netherlands
```

```
## country.New.Zealand
## country.Nigeria
## country.Norway
## country.Other
## country.Pakistan
## country.Peru
## country.Philippines
## country.Poland
## country.Portugal
## country.Republic.of.Korea
## country.Romania
## country.Russia
## country.Singapore
## country.South.Africa
## country.South.Korea
## country.Spain
## country.Sweden
## country.Switzerland
## country.Thailand
## country.Tunisia
## country.Turkey
## country.Ukraine
## country.United.Kingdom.of.Great.Britain.and.Northern.Ireland
## country.United.States.of.America
## country.Viet.Nam
## job_role.Business.Analyst
## job_role.Chief.Officer
## job_role.Consultant
## job_role.Data.Analyst
## job_role.Data.Engineer
## job_role.Data.Journalist
## job_role.Data.Scientist
## job_role.DBA.Database.Engineer
## job_role.Developer.Advocate
## job_role.Manager
## job_role.Marketing.Analyst
## job_role.Other
## job_role.Principal.Investigator
## job_role.Product.Project.Manager
## job_role.Research.Assistant
## job_role.Research.Scientist
## job_role.Salesperson
## job_role.Software.Engineer
## job_role.Statistician
## job_role.Student
## industry.Academics.Education
## industry.Accounting.Finance
## industry.Broadcasting.Communications
## industry.Computers.Technology
## industry.Energy.Mining
## industry.Government.Public.Service
## industry.Hospitality.Entertainment.Sports
## industry.I.am.a.student
## industry.Insurance.Risk.Assessment
```

```
## industry.Manufacturing.Fabrication
## industry.Marketing.CRM
## industry.Medical.Pharmaceutical
## industry.Military.Security.Defense
## industry.Non.profit.Service
## industry.Online.Business.Internet.based.Sales
## industry.Online.Service.Internet.based.Services
## industry.Other
## industry.Retail.Sales
## industry.Shipping.Transportation
## ML_atwork.I.do.not.know
## ML_atwork.No..we.do.not.use.ML.methods.
## ML_atwork.We.are.exploring.ML.methods..and.may.one.day.put.a.model.into.production.
## ML_atwork.We.have.well.established.ML.methods..i.e...models.in.production.for.more.than.2.years.
## ML_atwork.We.recently.started.using.ML.methods..i.e...models.in.production.for.less.than.2.years.
## ML_atwork.We.use.ML.methods.for.generating.insights..but.do.not.put.working.models.into.production.
## percent_actively.coding.0..of.my.time
## percent_actively.coding.1..to.25..of.my.time
## percent_actively.coding.100..of.my.time
## percent_actively.coding.25..to.49..of.my.time
## percent_actively.coding.50..to.74..of.my.time
## percent_actively.coding.75..to.99..of.my.time
## How.long.have.you.been.writing.code.to.analyze.data....1.year
## How.long.have.you.been.writing.code.to.analyze.data..1.2.years
## How.long.have.you.been.writing.code.to.analyze.data..10.20.years
## How.long.have.you.been.writing.code.to.analyze.data..20.30.years
## How.long.have.you.been.writing.code.to.analyze.data..3.5.years
## How.long.have.you.been.writing.code.to.analyze.data..30.40.years
## How.long.have.you.been.writing.code.to.analyze.data..40..years
## How.long.have.you.been.writing.code.to.analyze.data..5.10.years
## How.long.have.you.been.writing.code.to.analyze.data..I.have.never.written.code.and.I.do.not.want.to.l
## How.long.have.you.been.writing.code.to.analyze.data..I.have.never.written.code.but.I.want.to.learn
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school.....1.year
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...1.2.years
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...10.15.years
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...2.3.years
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...20..years
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...3.4.years
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...4.5.years
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...5.10.years
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...I.have.never.studi
## For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...I.have.never.studi
```

```r
varImpPlot(rf_model, main = "Variable Importance for Wage Prediction (RF)")
```



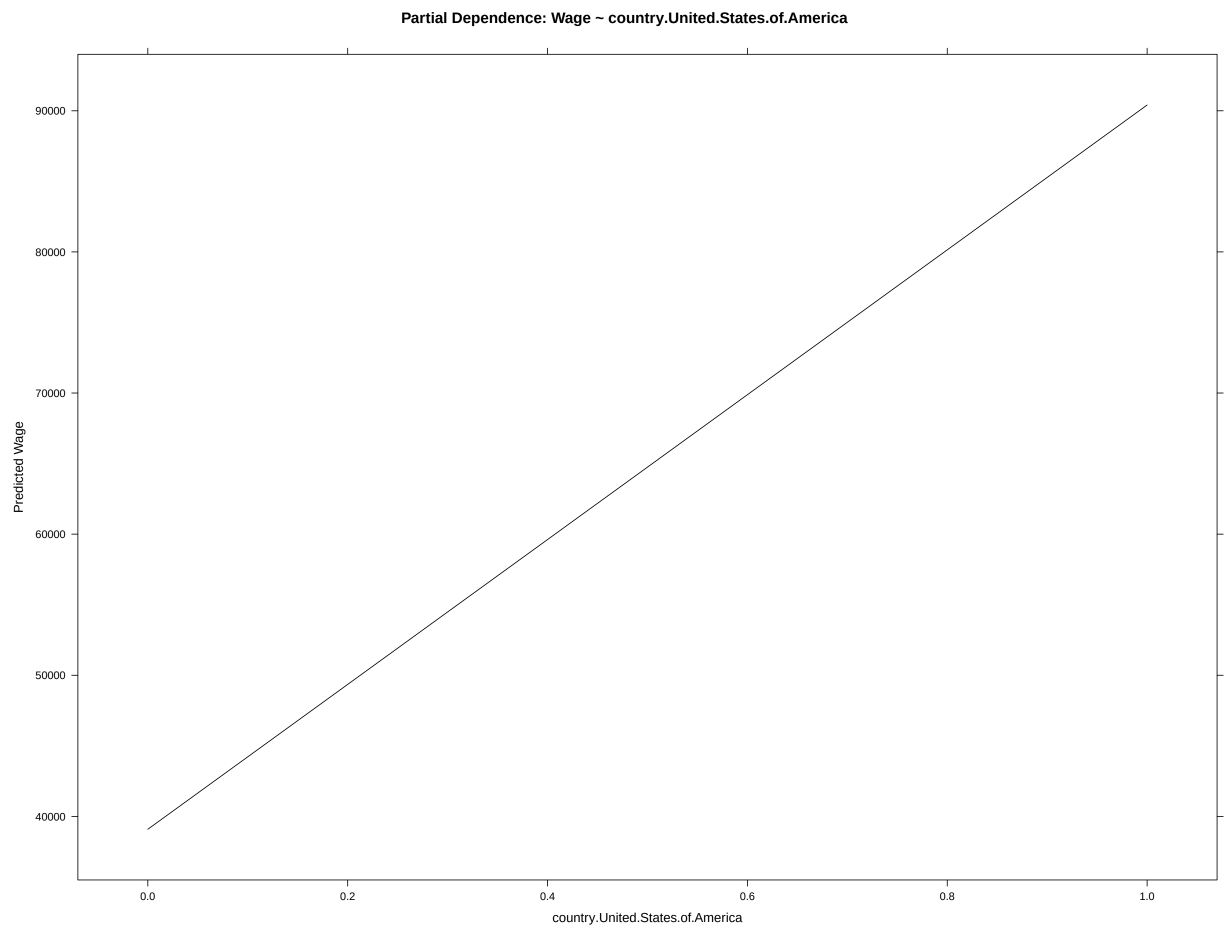

```r
cat("Top 5 RF features:\n")
```

```
## Top 5 RF features:
```

```r
print(head(sort(importance(rf_model)[,1], decreasing = TRUE), 5))
```

```
## country.United.States.of.America                country.Australia
##                        134.94110                         38.75356
##                 job_role.Student      industry.Academics.Education
##                         30.84257                         29.07535
##              country.Switzerland
##                         27.46158
```

```r
# Partial dependence plot for United states
pdp_experience <- partial(rf_model, pred.var = "country.United.States.of.America", train = train_data)
plotPartial(pdp_experience, main = "Partial Dependence: Wage ~ country.United.States.of.America",
            xlab = "country.United.States.of.America", ylab = "Predicted Wage")
```



11 AutoML VS our own Random Forest model

```
# Performance of AutoML model
perf_best_rmse <- h2o.performance(best_model, valid)
aml_rmse <- h2o.rmse(perf_best_rmse)
aml_mae  <- h2o.mae(perf_best_rmse)
aml_r2   <- h2o.r2(perf_best_rmse)

# Performance of Random Forest (train/test split)
rf_model <- randomForest(wage ~ ., data = train_data, importance = TRUE)
rf_pred <- predict(rf_model, test_data)
rf_rmse <- sqrt(mean((test_data$wage - rf_pred)^2))
rf_mae  <- mean(abs(test_data$wage - rf_pred))
rf_r2   <- cor(test_data$wage, rf_pred)^2

# Results summary table
results_table <- data.frame(
  Model = c("Random Forest", "H2O AutoML"),
  RMSE = c(rf_rmse, aml_rmse),
  MAE  = c(rf_mae, aml_mae),
  R2   = c(rf_r2, aml_r2)
```

```
)
print(results_table)
```

```
##            Model     RMSE      MAE        R2
## 1 Random Forest 44012.25 23502.59 0.5048184
## 2    H2O AutoML 40345.91 23030.66 0.5644355
```

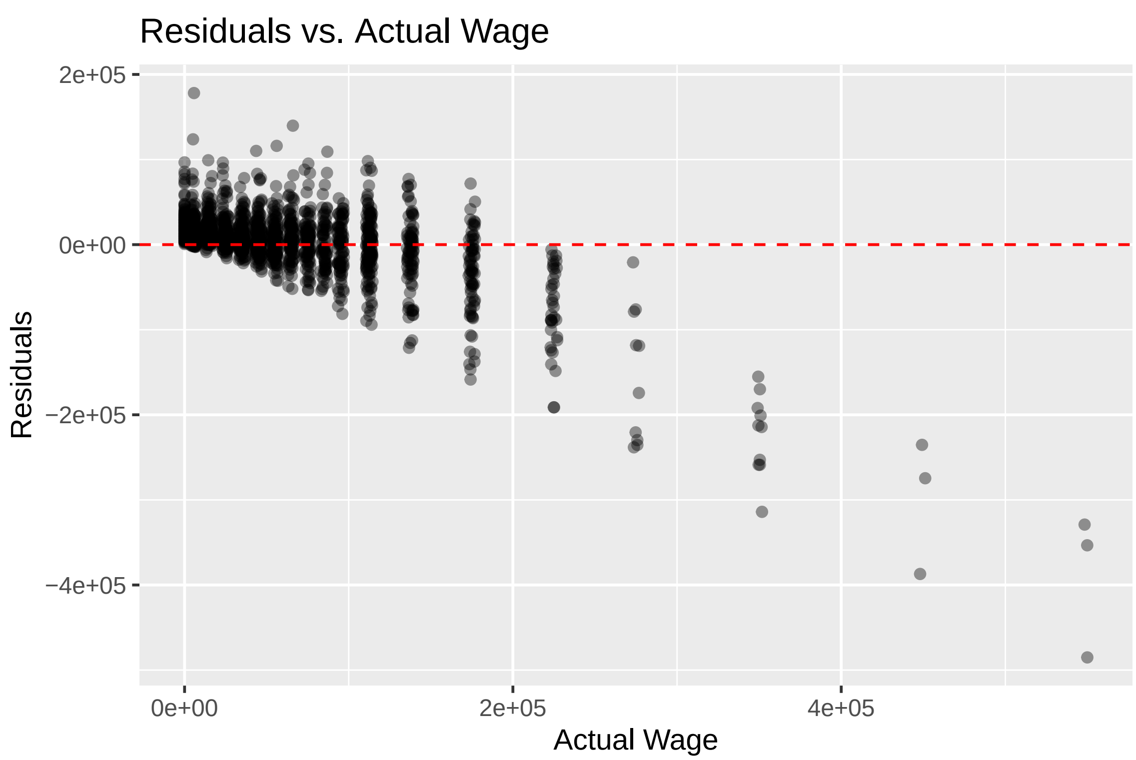
12 Plot residuals and predicted vs actual values:

We visualize how well the model predicted wages using residual and prediction plots.

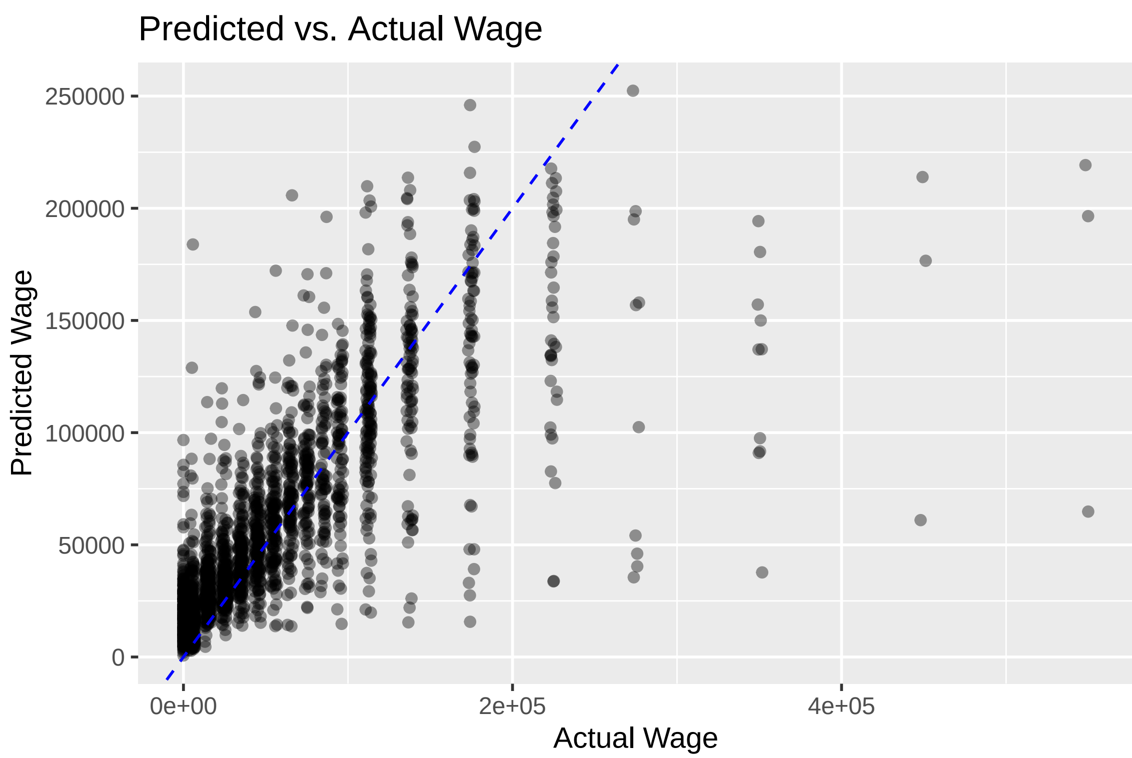
```
# AutoML predictions
pred_best_rmse <- h2o.predict(best_model, valid)
```

```
##   |                                                                              |
```

```
pred_df <- as.data.frame(h2o.cbind(pred_best_rmse, valid$wage))
colnames(pred_df) <- c("predicted", "actual")

# Residuals plot
ggplot(pred_df, aes(x = actual, y = predicted - actual)) +
  geom_point(alpha = 0.4) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs. Actual Wage", x = "Actual Wage", y = "Residuals")
```



```r
# Predicted vs actual
ggplot(pred_df, aes(x = actual, y = predicted)) +
  geom_point(alpha = 0.4) +
  geom_abline(slope = 1, intercept = 0, color = "blue", linetype = "dashed") +
  labs(title = "Predicted vs. Actual Wage", x = "Actual Wage", y = "Predicted Wage")
```



13 Real-world Application: Predicting Team Member Wages

We use the trained model to predict the wages of team members based on their profile.

```r
team_raw <- data.frame(
  age = factor(c("22-24", "22-24", "22-24", "25-29"), levels = levels(df$age)),
  years_experience = factor(c("0-1", "0-1", "3-4", "5-11"), levels = levels(df$years_experience)),
  education = factor(c("Bachelor's degree", "Bachelor's degree", "Bachelor's degree", "Bachelor's degre
  gender = factor(c("Male", "Male", "Female", "Female"), levels = levels(df$gender)),
  country = factor(c("Denmark", "Belgium", "Switzerland", "Switzerland"),
                levels = levels(df$country)),
  job_role = factor(c("Data Scientist", "Student", "Other", "Other"), levels = levels(df$job_role)),
  industry = factor(c("Computers/Technology", "Computers/Technology", "Computers/Technology", "Computer
  ML_atwork = factor(c("We have well established ML methods (i.e., models in production for more than 2
  percent_actively.coding = factor(c("75% to 99% of my time", "25% to 49% of my time", "0% of my time",
  For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school.. = factor(c("3-4 yea
  How.long.have.you.been.writing.code.to.analyze.data. = factor(c("3-5 years", "3-5 years", "< 1 year",

)
```

```r
team_dummy <- predict(dummy_model, newdata = team_raw)
team_matrix <- data.frame(team_raw, team_dummy)
# Convert to H2OFrame
team_h2o <- as.h2o(team_matrix)
```

```
##   |                                                                              |
```

```r
# Predict and assign
team_matrix$predicted_wage <- as.vector(predict(GBM_rsme, newdata = team_h2o))
```

```
##   |                                                                              |
```

```r
team_matrix[, c("age", "years_experience", "education", "gender", "country", "job_role", "industry","ML_
```

```
##     age years_experience         education gender     country       job_role
## 1 22-24              0-1 Bachelor's degree   Male     Denmark Data Scientist
## 2 22-24              0-1 Bachelor's degree   Male     Belgium        Student
## 3 22-24              3-4 Bachelor's degree Female Switzerland          Other
## 4 25-29             5-11 Bachelor's degree Female Switzerland          Other
##              industry
## 1 Computers/Technology
## 2 Computers/Technology
## 3 Computers/Technology
## 4 Computers/Technology
##                                                                              ML_atwork
## 1    We have well established ML methods (i.e., models in production for more than 2 years)
## 2   We recently started using ML methods (i.e., models in production for less than 2 years)
## 3                                                                           I do not know
## 4 We use ML methods for generating insights (but do not put working models into production)
##   percent_actively.coding
## 1   75% to 99% of my time
## 2   25% to 49% of my time
## 3           0% of my time
## 4           0% of my time
##   For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school..
## 1                                                                         3-4 years
## 2                                                                         3-4 years
## 3             I have never studied machine learning but plan to learn in the future
## 4                                                                          < 1 year
##   How.long.have.you.been.writing.code.to.analyze.data. predicted_wage
## 1                                            3-5 years      48954.180
## 2                                            3-5 years       9732.364
## 3                                             < 1 year      66757.178
## 4                                           5-10 years      88222.351
```

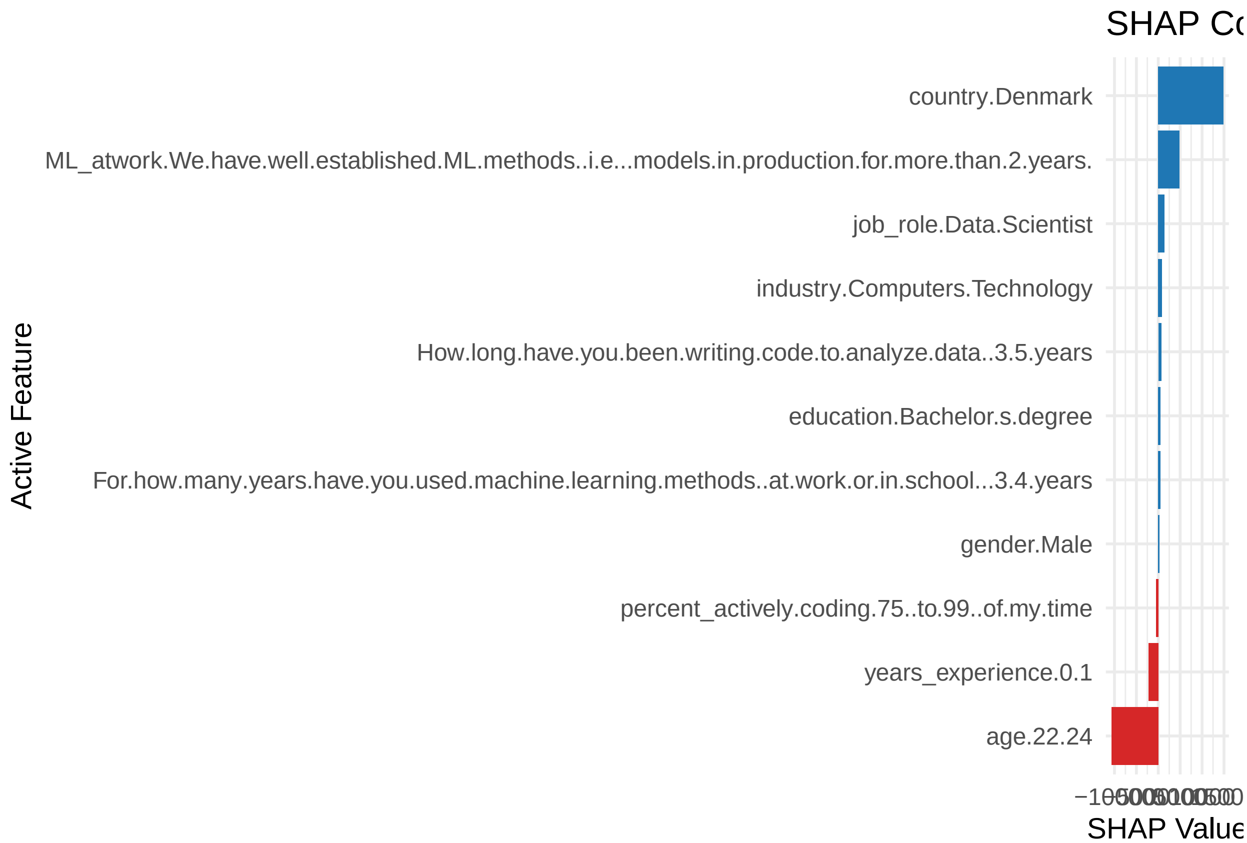
Our results: Predicted wages Row 1 = Ian: 73822.06 Row 2 = Piet: 26811.76 Row 3 = Rahel: 64331.60 Row 4 = Dana: 97305.01

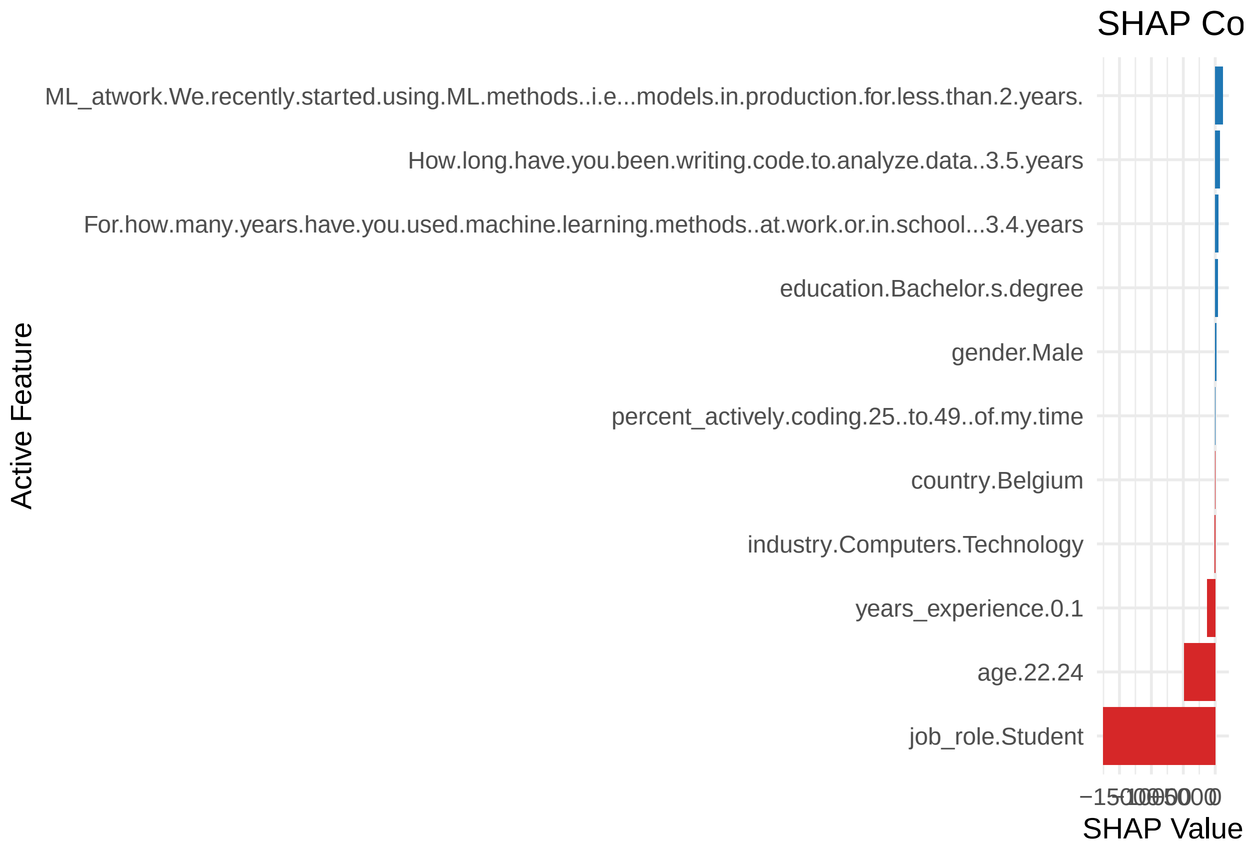
A shap plot per person to see why they earn the amount

```r
library(ggplot2)

# Get SHAP values and dummy features
shap_values <- h2o.predict_contributions(GBM_rsme, team_h2o)
```

```
##   |                                                                              |
```

```r
shap_df <- as.data.frame(shap_values)
team_dummy_df <- as.data.frame(team_h2o)

# Loop through each individual
for (i in 1:4) {
  individual_shap <- shap_df[i, ]
  individual_input <- team_dummy_df[i, ]

  # Remove BiasTerm
  individual_shap <- individual_shap[, !(names(individual_shap) == "BiasTerm")]
  individual_input <- individual_input[, !(names(individual_input) == "BiasTerm")]

  # Find active features (equal to 1)
  active_feature_names <- names(individual_input)[which(individual_input == 1)]

  # Filter SHAP values by active feature names
  shap_filtered <- data.frame(
    Feature = active_feature_names,
    SHAP_value = unlist(individual_shap[active_feature_names])
  )

  # Sort by absolute SHAP value
  shap_filtered <- shap_filtered[order(abs(shap_filtered$SHAP_value), decreasing = TRUE), ]

  # Plot
  p <- ggplot(shap_filtered, aes(x = reorder(Feature, SHAP_value), y = SHAP_value, fill = SHAP_value > 
    geom_col(show.legend = FALSE) +
    coord_flip() +
    labs(
      title = paste("SHAP Contributions (Active Features Only) - Individual", i),
      x = "Active Feature",
      y = "SHAP Value"
    ) +
    scale_fill_manual(values = c("TRUE" = "#1f77b4", "FALSE" = "#d62728")) +
    theme_minimal()


  print(p)
}
```


SHAP Co
Active Feature
country.Denmark
ML_atwork.We.have.well.established.ML.methods..i.e...models.in.production.for.more.than.2.years.
job_role.Data.Scientist
industry.Computers.Technology
How.long.have.you.been.writing.code.to.analyze.data..3.5.years
education.Bachelor.s.degree
For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...3.4.years
gender.Male
percent_actively.coding.75..to.99..of.my.time
years_experience.0.1
age.22.24
SHAP Value


SHAP Co
Active Feature
ML_atwork.We.recently.started.using.ML.methods..i.e...models.in.production.for.less.than.2.years.
How.long.have.you.been.writing.code.to.analyze.data..3.5.years
For.how.many.years.have.you.used.machine.learning.methods..at.work.or.in.school...3.4.years
education.Bachelor.s.degree
gender.Male
percent_actively.coding.25..to.49..of.my.time
country.Belgium
industry.Computers.Technology
years_experience.0.1
age.22.24
job_role.Student
SHAP Value

co

education.

yea

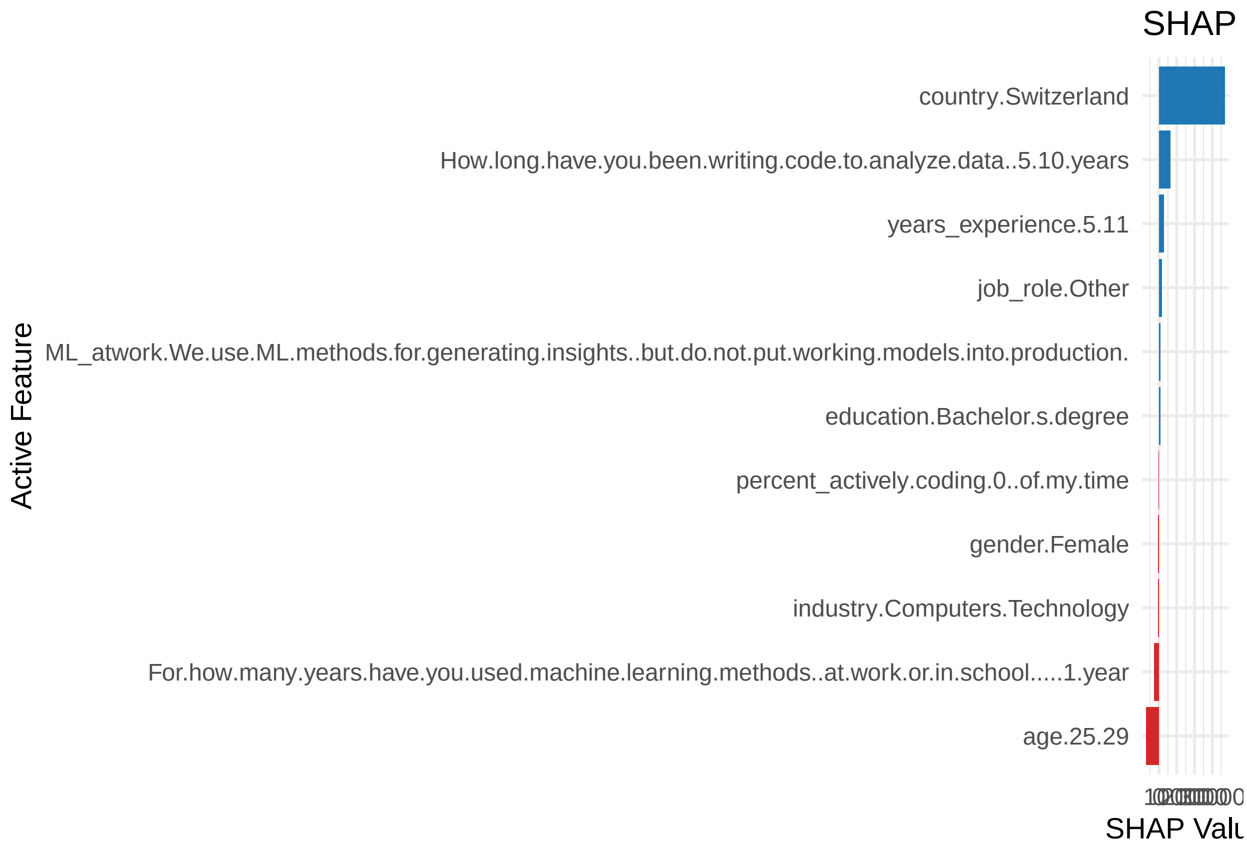
percent_actively.co

ears.have.you.used.machine.learning.methods..at.work.or.in.school...I.have.never.studied.machine.learning.but.plan.to

ML_atv

industry.Com

How.long.have.you.been.writing.code.to.ana



These shap plots should explain exactly why everyone got their wage but underneath we'll give a more indepth explanation:

The explanation, Ian, Piet, and, Rahel are all the same age, but Dana is older and as we could see from the partial dependency plot and feature importance plot, the older you are the higher your wage is, up untill 70 years old. Also years of experience plays a big role, Dana once again has the most followed by Rahel and with no experience we have Ian and Piet. For industry we have all said that we will work in the Computers/Technology industry, and male's get paid higher in this industry, boosting the wages of Ian and Piet just a small bit. As gender doesn't play that big of a factor, you can see this from the feature importance plot. On the other hand Country does play a big role, For Rahel and Dana, they plan to work in Switzelrand which has a very strong economy with an average pay that's way higher than Belgium, Ian is planning on working in Denmark which lays somewhere in between Switzerland and Belgium. Piet filled everything in as he was working full time except for job_role, here he said that he would be a student which severely impacted his wage as, job_role being student is high up the variable importance plot. This would explain why his pay would be relatively little compared to the others. Ian and Rahel are also kind of similar, although Rahel would work in Switzerland which should get paid more, Ian has more experience when it comes to Machine learning, and his job would be more coding focused. Resulting in a higher pay. Aslo Rahel's job_role is Other as the job which she would be doing is not in the list, if her specific job would be in the list she might end up with a higher wage than Ian. The same goes for Dan, her job_role is currently also Other, if her specific role was in the list to choose from she might end up with an even higher wage. Than you might say are Dana and Rahel not very similar and shouldn't they get payed about the same, as they are both female, in Switzerland, with the job_role other? No, there is still a more than 30.000 difference which could be due to like mentioned before, experience but also Dana has had between 5-10 years of writing code to analyze data.

14 Conlusion

In this project, we tried to predict wages based on an extensive array of features of a very large dataset, using both standard machine learning techniques and AutoML techniques. After careful data exploration, we identified several key features that have a very critical role in predicting wages, such as years of experience, industry, education, and some technical skills. We also overcame problems such as possible outliers—specifically young high earners—and discussed the use of sensitive variables such as gender, finally justifying our actions using statistical evidence.

Our investigation suggested that individual attributes alone do not determine wage outcomes; instead, it is the interaction of factors such as occupational function, industry, technical expertise, and coding experience that leads to predictive accuracy. Secondly, employing AutoML also circumvented tedious model selection and optimization by virtue of streamlined selection and adjustment, enabling us to acquire stable results without needing exhaustive hand trials.

All in all, this project speaks to the value of systematic, data-focused solutioning towards predicting wages. With the combination of domain knowledge, ethical concerns, and advanced machine learning resources, we were able to construct comprehensible and precise models. Future work would include enhancing outlier detection, incorporating more external sources of data, and validating more complex ensemble methods. However, our findings provide a solid foundation for future research on determining factors of wage disparities in the technology and data science sectors.

15 Save Model and Results

Optionally, you can persist your model and predictions for deployment or future reuse. As retraining the model takes a long time.

```r
save(rf_model, GBM_rsme, best_model, file = "wage_model.RData")
write.csv(team_matrix, file = "team_wage_predictions.csv", row.names = FALSE)
```

End of Report