BDA Project

Akseli Manninen, Niko Miller, and Santeri Löppönen

Contents

1. Project Description and Motivation

Studying the relationship between income and education has been the focus on many studies. The studies have concluded that there is a strong correlation between higher education and income (Card, 1999). In general, individuals with stronger education are more likely to be employed and earn a big salary compared to less educated people (Card, 1999). For that reason, education is described as an investment in human capital (Wolla & Sullivan, 2017).

This study examines this phenomenon from the perspective of people that have acquired their education from colleges of the United States. As the connection between education and income has been shown in the existing literature, this study strives to examine the associations between college related features and income level years after graduation. This project is not limited to only considering educational aspects but expands it to family backgrounds.

In this study, a Bayesian approach is taken to observe the bond between the educational and family related features and earnings. It is in our interest to find out, how accurately the selected features can predict future income for the students. Furthermore, finding a well-predictive features among the vast number of variables is pursed and evaluating the predictive performance of selected statistical models. As this study is conducted in a university environment by university students, the possible insight would be especially meaningful for the members of the group and peers.

2. Data and the analysis problem

We use the most recent institutional-level college scorecard data from the US Department of Education. The institutional-level dataset contains aggregate data for each educational institution and includes data on institutional characteristics, enrollment, student aid, costs and student outcomes. The dataset has over 6000 observations on more than 3000 variables.

We chose to use this dataset because we could use the data to answer interesting education-related questions in our project, and also because the dataset has a large number of observations and also a large number of variables. The large amount of variables means we would have a lot of flexibility when it came to modelling our data and there would be enough data to possibly make valid inferences due to many observations. Of course, the large amount of variables also posed challenges, as it was arguably slower and more burdensome to find the most relevant variables to use in our analysis. We were somewhat surprised by how fast the observation count started to shrink when we began cleaning the data, so in hindsight we should have maybe paid more attention to cleanliness, as this dataset had for example a lot of missing values

```
# read in data
data.joined.stan <- read.csv(file = "./Data/data.joined.stan.csv")
numerical.vars.data <- read.csv(file = "./Data/numerical.vars.data.csv")
categorical.vars.data <- read.csv(file = "./Data/categorical.vars.data.csv")
```

Feature selection

Phase 1 - Feature selection based on literature and domain knowledge

The initial data set had almost 3000 features and in that regards, the number of observations is relatively small. There are also a lot of missing values in the data set and some features are missing. For these reasons, there was a need to prune features.

The used process of feature selection consisted of two phases: In the first phase, a subset of features was select based on the features used in the existing literature and using domain knowledge. From the potential features, only those having enough data were included in the subset and others were discarded.

The selected features were:

	Name	*Type*	*Description*
1	$SATVRMID$	*numerical*	$Midpoint\ of\ SAT\ scores\ at\ the\ institution\ (critical\ reading)$
2	$SATMTMID$	*numerical*	$Midpoint\ of SAT scores at the institution (math)$
3	$SATWRMID$	*numerical*	$Midpoint\ of\ SAT\ scores\ at\ the\ institution\ (writing)$
4	MD_FAMINC	*numerical*	$Median\ family\ income$
5	AGE_ENTRY	*numerical*	$Average\ age\ of\ entry$
6	$FEMALE$	*numerical*	$Share\ of\ female\ students$
7	$FIRST\_GEN$	*numerical*	$Share\ of\ first-generation students$
8	PCT_WHITE	*numerical*	$Percent\ of\ the\ population\ from\ students'\ zip\ codes\ that\ is\ White$
9	$DEBT\_MDN\_SUPP$	*numerical*	$Median\ debt,\ suppressed\ for\ n = 30$
10	$C150\_4$	*numerical*	$Completion\ rate\ for\ first-time,\ full-tim\ students$
11	$COSTT4\_A$	*numerical*	$Average\ cost\ of\ attendance\ (academic\ year\ institutions)$
12	$POVERTY\_RATE$	*numerical*	$Poverty\ rate$
13	$UNEMP\_RATE$	*numerical*	$Unemployment\ rate$
14	$MARRIED$	*numerical*	$Share\ of\ married students$
15	$VETERAN$	*numerical*	$Share\ of\ veteran students$
16	$LOCALE$	*categorical*	$Locale\ of\ institution$
17	$CCBASIC$	*categoriacal*	$Carnegie\ Classification--basic$
18	$CONTROL$	*categorical*	$Control\ of\ institution$
19	$MD\_EARN\_WNE\_P10$	*numerical*	$Median\ earnings\ of\ students\ 10\ years\ after\ entry$

Phase 2 - Feature selection wiht correlation and visual dependency

In the second phase of feature selection, a subset of features was selected from the 18 variables of the first phase. The correlations between the features were examined as well as their associations to the dependent variable.

Numerical variables

SAT scores were combined as one variable, because they were correlated and viewed as one entity. However, writing SAT scores had too few observations, due to discontinued tracking, so the variable SAT_ALL was formed by summing the math and critical thinking SAT scores. SAT scores were included because the correlation was high with the dependent variable.

For the rest of the numerical variables, if the correlation between a feature and the dependent variable was low and there was not observable dependency between the two in the scatter plot, the feature was excluded. Also, if multiple independent variables were highly correlated, for the sake of simplicity only one of them was selected to the model, based on the highest correlation and dependency with the dependent variable.

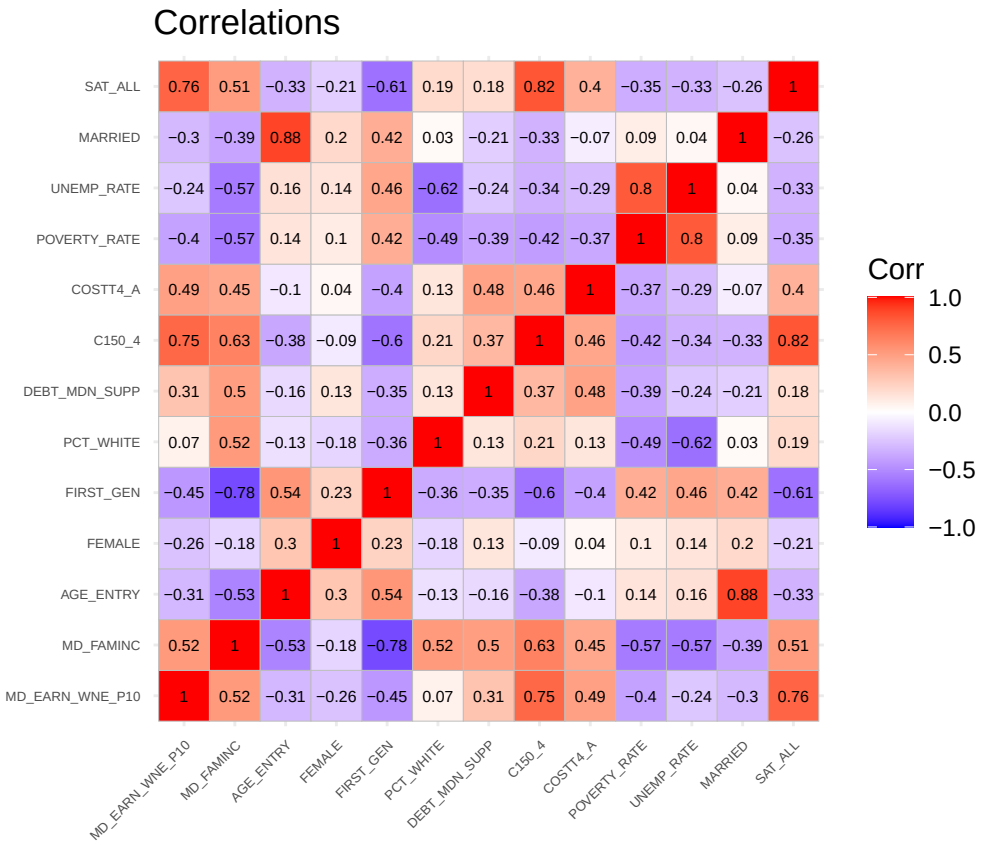
The selected numerical variables were: SAT_ALL, MD_FAMINIC, AGE_ENTRY, COSTT4_A, and POVERTY_RATE.

Visualizing correlations and data points with the dependent variable.

Pearson's Correlation

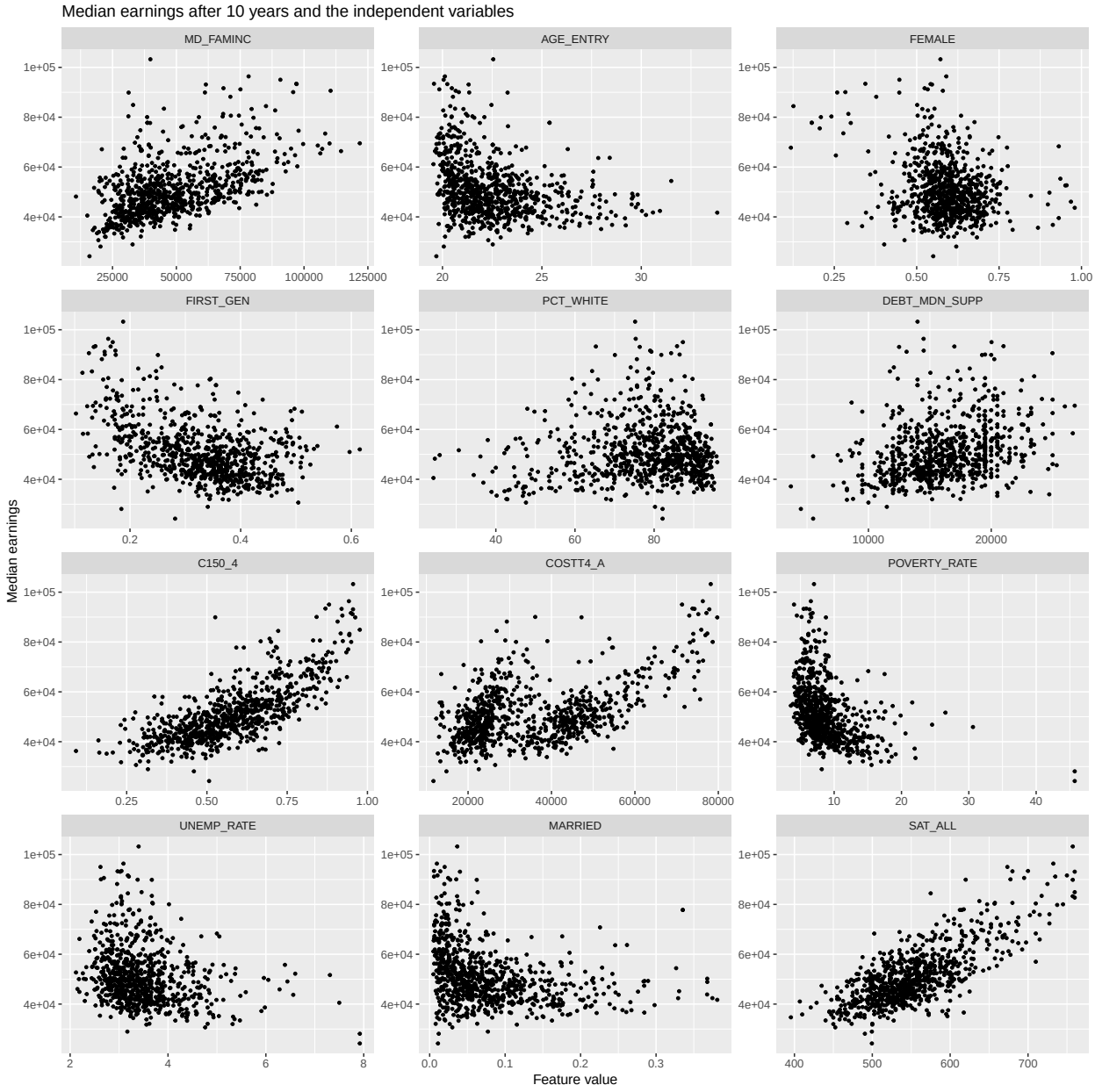
```
# correlation plot
ggcorrplot(cor(numerical.vars.data),
```

```
          lab = TRUE,
          lab_size = 2,
          title = "Correlations",
          tl.cex = 5)
```



Dependent variable vs. feature scatter plots

```
melted.numerical.vars.data <- melt(numerical.vars.data, id.vars = "MD_EARN_WNE_P10")
ggplot(melted.numerical.vars.data, aes(x = value, y = MD_EARN_WNE_P10)) +
  facet_wrap(~variable, scales = "free", ncol = 3) +
  geom_point(shape=20, color="black") +
  ggtitle("Median earnings after 10 years and the independent variables") +
  xlab("Feature value") + ylab("Median earnings")
```



Categorical variables

The categorical variables were observed with box plots to see if there were differences between the categories and the dependent variable. Based on an analysis on the variable LOCALE, a new binary variable URBAN was generated which represents, where value 1 represents city-like area which includes categories Large City, Mid-Size City, Urban Fringe of a Large City and Urban Fringe of a Mid-Size City of the LOCALE variable from the dataset.

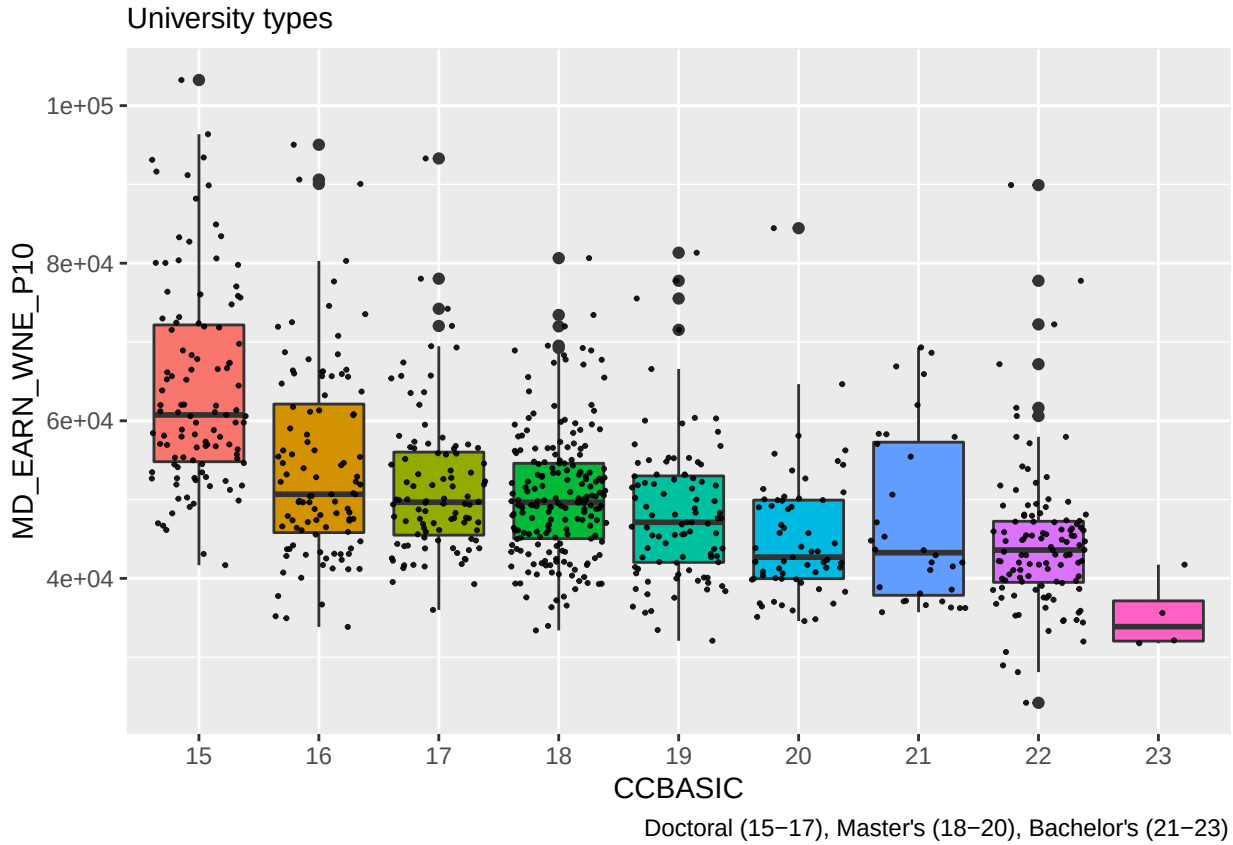
The categorical variable CCBASIC which represented the Carnegie Classification was divided into two binary MASTER and DOCTORL. These variables represent if the college is classified as Master's college and university or Doctoral's college and university. If a college does not belong to either of those, it is a Bachelor's college and university. Other special focus colleges and universities were discarded from the dataset to keep the model more simple.

The categorical variable CONTROL, which had three classes Public, Private and Private, Nonprofit and Proprietary was modified into binary variable Private representing if the school is private or public school. There were only few Proprietary observations, so those were discarded.

The selected categorical variables were: URBAN, DOCTORAL, MASTER, PRIVATE.

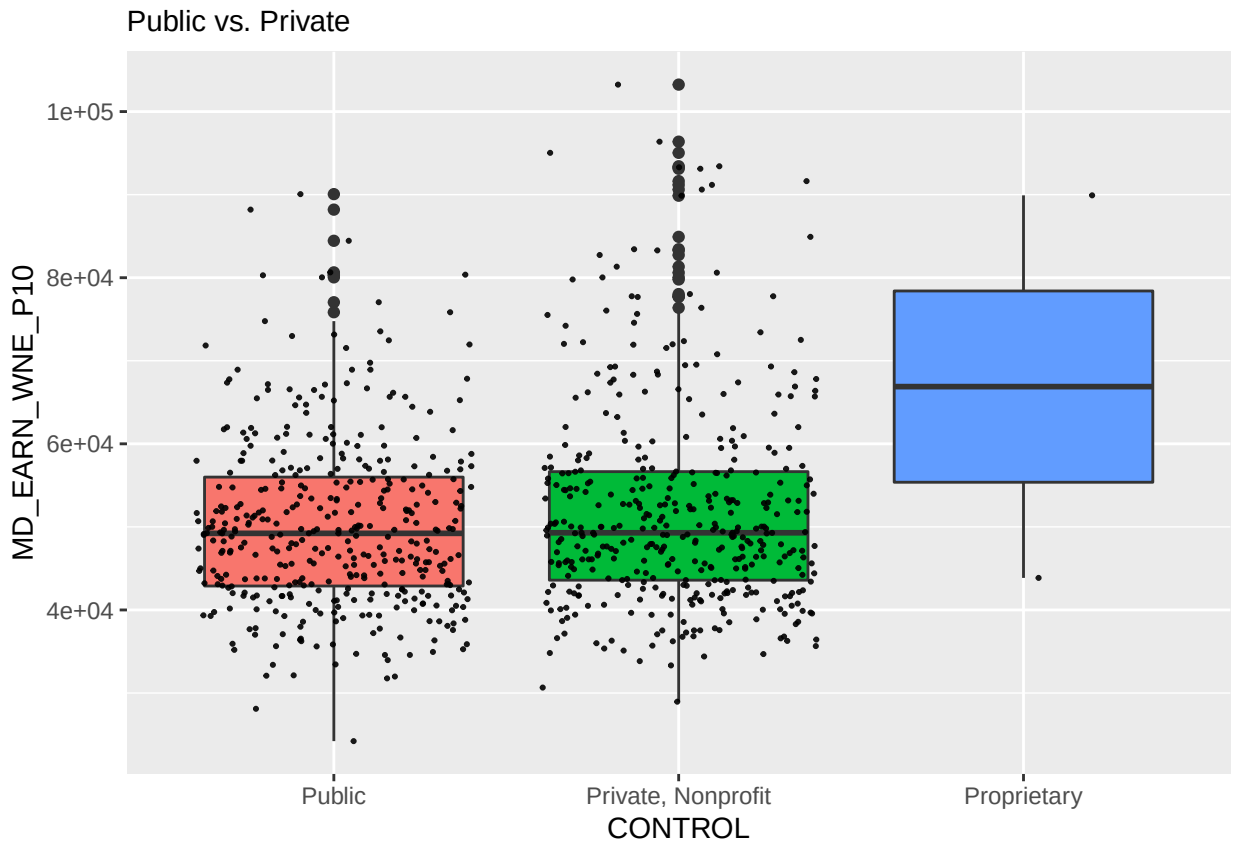
```
# categorical variable box plots
melted.categorial <- melt(categorical.vars.data, id.vars = "MD_EARN_WNE_P10")
melted.categorial.ccbasic <- melted.categorial %>% as_tibble() %>% filter(variable=="CCBASIC")
melted.categorial.control <- melted.categorial %>% as_tibble() %>% filter(variable=="CONTROL")
melted.categorial.locale <- melted.categorial %>% as_tibble() %>% filter(variable=="LOCALE")

ggplot(melted.categorial.ccbasic, aes(x=value %>% as.factor(), y=MD_EARN_WNE_P10, fill=value %>% factor
  geom_boxplot() +
  geom_jitter(color="black", size=0.4, alpha=0.9) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("University types") +
  xlab("CCBASIC") +
  labs(caption = "Doctoral (15-17), Master's (18-20), Bachelor's (21-23)")
```

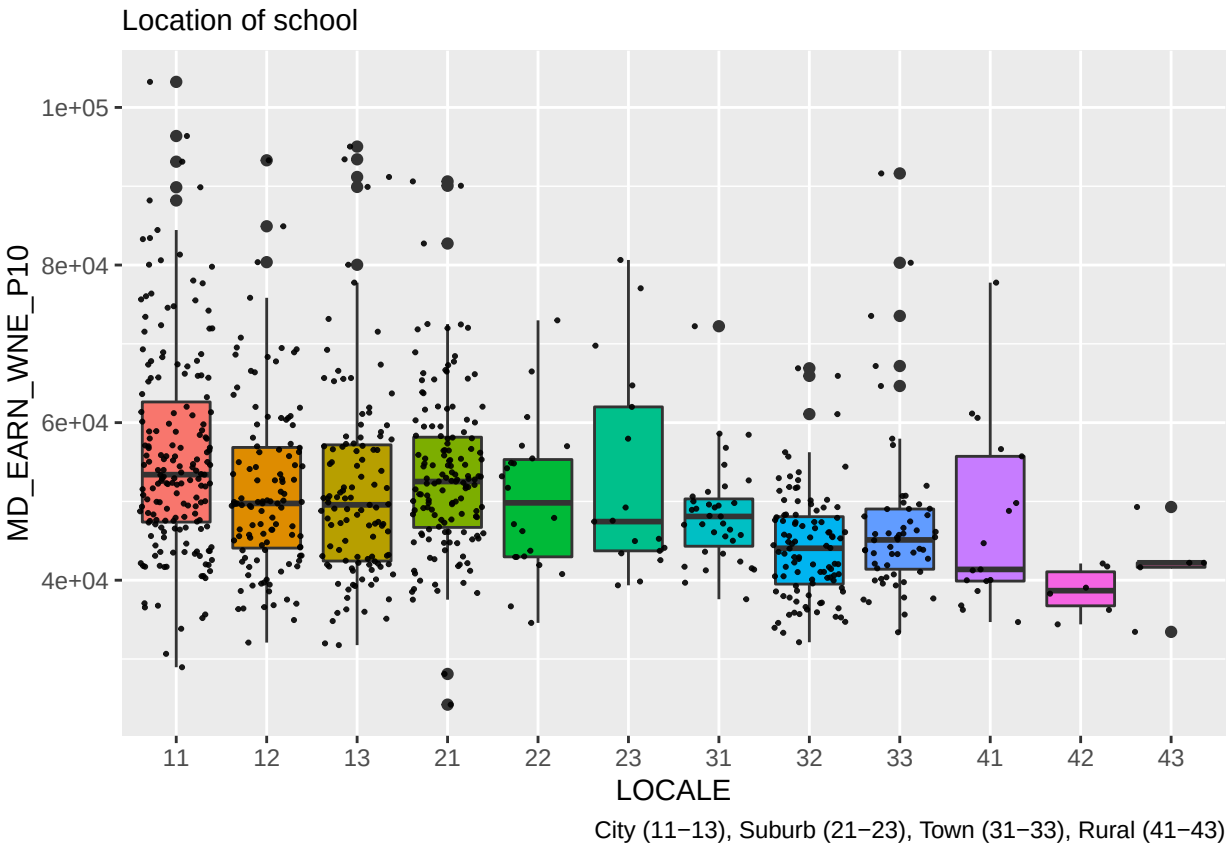



```
ggplot(melted.categorial.control, aes(x=value %>% as.factor(), y=MD_EARN_WNE_P10, fill=value %>% as.fac
  geom_boxplot() +
  geom_jitter(color="black", size=0.4, alpha=0.9) +
```

```
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("Public vs. Private") +
  xlab("CONTROL") +
  scale_x_discrete(labels = c("Public","Private, Nonprofit","Proprietary"))
```




```
ggplot(melted.categorial.locale, aes(x=value %>% as.factor(), y=MD_EARN_WNE_P10, fill=value %>% as.facto
  geom_boxplot() +
  geom_jitter(color="black", size=0.4, alpha=0.9) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("Location of school") +
  xlab("LOCALE") +
  labs(caption = "City (11-13), Suburb (21-23), Town (31-33), Rural (41-43)")
```



Phase 3 - Feature selection with Stepwise regression

In the third phase, the remaining variables were used with Stepwise regression, to test which subset of features perform the best, and whether the received coefficients are reasonable, and if there are signs of overfitting.

The Stepwise regression suggested using all of the variables, except CITY and DOCTORAL. The stepwise regression can be seen in appendix 1.

The final features are listed in the table below:

	Name	*Data type*	*Description*
1	SAT_ALL	*float*	*Midpoint of SAT scores at the institution (critical reading , math)*
2	MD_FAMINC	*float*	*Median family income*
3	AGE_ENTRY	*float*	*Average age of entry*
4	$COSTT4\_A$	*float*	*Average cost of attendance (academic year institutions)*
5	$POVERTY\_RATE$	*float*	*Poverty rate*
6	$PRIVATE$	*binary*	*Carnegie Classification − −basic*
7	$MASTER$	*binary*	*Control of institution*
8	$MD\_EARN\_WNE\_P10$	*float*	*Median earnings of students 10 years after entry*

3. Description of the models

Description of the separate model

Add LaTeX!

Description of the pooled model

Add LaTeX!

Description of the hierarchical model

Add LaTeX!

Description of the linear model

4. Informative or weakly informative priors, and justification of their choices.

The selected priors are mostly weakly informative, as we do not posses enough information about the dependency between the independent variables and the dependent variable. When selecting the priors, proper distribution type was considered for each variable. What comes to the standard deviations, they were selected based on the magnitude of absolute values of the variables, and considering what kind of impact they could have for income, with somewhat exaggerated estimates to avoid limiting the values too much.

SAT_ALL ~ Normal(43, 500)

Justification: Weakly informative prior as we don't posses enough information on the dependency, although it would be intuitive that success in the SAT exam would be associated with higher income. For that reason, the standard deviation is set to be high, and the variable is not limited to be greater than one.

The mean is calculated with the following formula: Median individual income in the United States / Average SAT Score (math and writing).

The median individual income in the US was approximately 31 000 in 2020 (Data Commons, 2020). The average SAT score in the US in 2020 were 523 in Math and 582 in Evidence-Based Reading and Writing (Number2, 2020).

mu = 31 000 / (523 + 528/2) = 43.2...

MD_FAMINIC ~ Normal(0, 100)

Justification: Weakly informative prior, is selected as we don't posses enough information on the dependency. The absolute values are in general high, (for example compared to AGE_OF_ENTRY) and thus the standard deviation is set lower for this variable.However, there could be cases where MD_FAMINIC is low, due to for example unemployment, so the standard deviation is still set considerably high.

AGE_ENTRY ~ Normal(0, 2500)

Justification: Weakly informative prior, is selected as we don't posses enough information on the dependency. As the age of entry could be somewhere in the range of 15 - 50 without considering outliers, a few dozen years difference could have dramatic changes in income either way, the standard deviation is set high.

COSTT4_A ~ Normal(0, 500)

Justification: Weakly informative prior, is selected as we don't posses enough information on the dependency. The average cost entry per academic year is most likely lower than median family income, but higher than average age of entry and thus the standard deviation is set between the standard deviations of those.

POVERTY_RATE ~ Normal(0, 2500)

Justification: Weakly informative prior, is selected as we don't posses enough information on the dependency. The possible values are between 0 and 100. There could be situations were poverty rate in an area is really low, for instance 0.5%. For that reason, the standard deviation in the prior is set high to enable possibly high weight for a small value.

MASTER ~ Normal(0, 2500)

PRIVATE ~ Normal(0, 2500)

Justification: For MASTER and PRIVATE weakly informative prior, is selected as we don't posses enough information on the dependency with the dependent variable. Because the values are always either 0 and 1, the standard deviation is set high.

5. Stan

Pooled model

```
pooled.model <- cmdstan_model(stan_file = "./Stan/pooled.stan")

pooled.model.data <- list(N = nrow(data.joined.stan),
                          y = data.joined.stan$MD_EARN_WNE_P10,
                          SAT_ALL = data.joined.stan$SAT_ALL,
                          MD_FAMINIC = data.joined.stan$MD_FAMINC,
                          AGE_ENTRY = data.joined.stan$AGE_ENTRY,
                          COSTT4_A = data.joined.stan$COSTT4_A,
                          POVERTY_RATE = data.joined.stan$POVERTY_RATE,
                          MASTER = data.joined.stan$MASTER,
                          PRIVATE = data.joined.stan$PRIVATE)


pooled.fit <- pooled.model$sample(data = pooled.model.data, seed = 1234, refresh = 1e3)

pooled.fit
```

Separate model

Lorem Ipsum

Hierarchical model

6. How to the Stan model was run, that is, what options were used.

This is also more clear as combination of textual explanation and the actual code line.

7. Convergence diagnostics (R^, ESS, divergences) and what was done if the convergence was not good with the first try.

This should be reported for all models.

Pooled model

```
params <- pooled.model$variables()$parameters %>% names()
rhat.df <- tibble()
for (param in params) {
  rhats <- extract_variable_matrix(pooled.fit$draws(), variable = param) %>% apply(2, rhat)
  row <- tibble("Parameter" = param,
                "Chain 1" = rhats[1],
                "Chain 2" = rhats[2],
                "Chain 3" = rhats[3],
                "Chain 4" = rhats[4])
  rhat.df <- rbind(rhat.df, row)
}
rhat.df
```

8. Posterior predictive checks and what was done to improve the model. This should be reported for all models.

9. Optional/Bonus: Predictive performance assessment if applicable (e.g. classification accuracy) and evaluation

of practical usefulness of the accuracy. This should be reported for all models as well.

10. Sensitivity analysis with respect to prior choices (i.e. checking whether the result changes a lot if prior

is changed). This should be reported for all models.

11. Discussion of issues and potential improvements.

- Correlation does not mean causality
- Ethics selecting features
-

12. Conclusion what was learned from the data analysis.

13. Self-reflection of what the group learned while making the project.

14. References

Insert bibliography here Card, D. (1999). THE CAUSAL EFFECT OF EDUCATION ON EARNINGS. Wolla, S. A., & Sullivan, J. (2017). Education, Income, and Wealth. https://fred.stlouisfed.org/graph/?g=7yKu.

Data Commons. (2020). Gross domestic product per capita in United States of America [Graph]. Retrieved from https://datacommons.org/place/country/USA?utm_medium=explore&mprop=income&popt=Person&cpv=age%2CYears15Onwards&hl=en

Number2. (2020). Average SAT Score [Blog Post]. Retrieved from: https://www.number2.com/average-sat-score/

15. Appendices

Appendix 1 - Stepwise regression

```
# PRELIMINARY ANALYSIS ----
# preliminary model with all numerical vars (not yet categorical)
# MODELING ----

data.joined.model <- data.joined.dropna %>%
  mutate(URBAN = case_when(LOCALE %in% c(seq(11,13), seq(21,23)) ~ 1,
                           TRUE ~ 0),
         PRIVATE = case_when(CONTROL %in% c(2,3) ~ 1,
                             TRUE ~ 0),
         DOCTORAL = case_when(CCBASIC %in% seq(15,17) ~ 1,
                              TRUE ~ 0),
         MASTER = case_when(CCBASIC %in% seq(18,20) ~ 1,
                            TRUE ~ 0)
  )


numerical.vars.model <- c("MD_EARN_WNE_P10", "SAT_ALL", "MD_FAMINC", "AGE_ENTRY", "COSTT4_A", "POVERTY_
categorical.vars.model <- c("URBAN", "PRIVATE", "DOCTORAL", "MASTER")

# data with REGION identifier for STAN
data.joined.stan <- data.joined.model %>%
  select(REGION, numerical.vars.model, categorical.vars.model)

# data for linear regression model in R
data.joined.model <- data.joined.stan %>%
    select(-REGION)

# baseline model
```

```r
model <- lm(MD_EARN_WNE_P10 ~ ., data = data.joined.model)
summary(model)

# step wise regression implied "best" model in terms of AIC
step(model, direction = "backward")
stepwise.model <- lm(formula = MD_EARN_WNE_P10 ~ SAT_ALL + MD_FAMINC + COSTT4_A + POVERTY_RATE + URBAN +
summary(stepwise.model)
```