BILLIONAIRES: THE BOTTOM LINE

Courtney Hodge, Charlie Perez, Will Peritz,
Bereket Tafesse, Hannah Valenty

INTRODUCTION

The Billionaires Dataset Motive

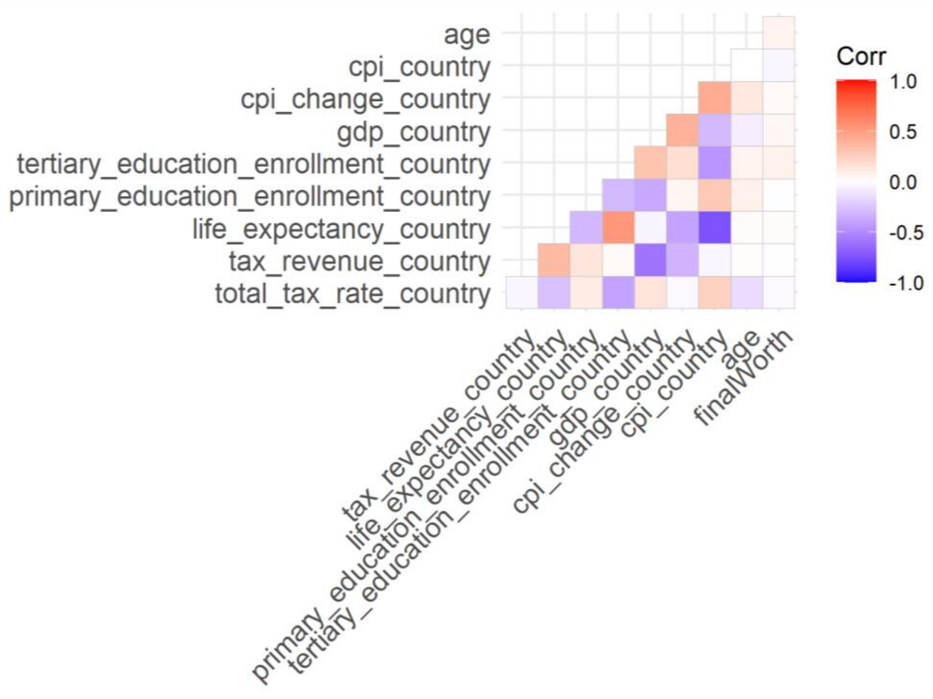
- Though many of the billionaires in our dataset are not a popular celebrity, their wealth raises them to a higher status in the public eye
- 81 billionaires combined have more wealth than 50% of the world combined
    - https://www.globalcitizen.org/en/content/wealth-inequality-oxfam-billionaires-elon-musk/
- Understanding how the bulk of wealth is distributed by billionaires across countries is necessary to understand the country's economic state
    - Are Americans wealthy or does the U.S. just have a lot of billionaires?
- Wealth can be self-made, but often is inherited from the previous generation

FORBES REAL-TIME BILLIONAIRES LIST

https://www.forbes.com/real-time-billionaires/#54c6f3cf3d78

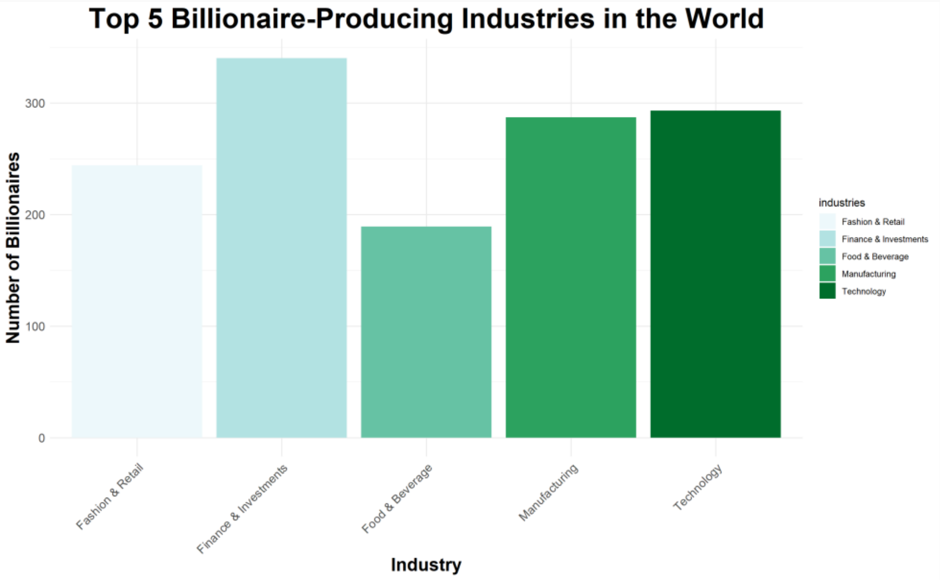
Variable Explanations	
age	Billionaire's age
finalWorth	Net worth of the billionaire (in millions)
_country	Given metric in the billionaire's country
cpi	Consumer Price Index
selfMade	Wether the billionaire is self-made or if the wealth is inherited
industries	What industry the billionaire's wealth is from

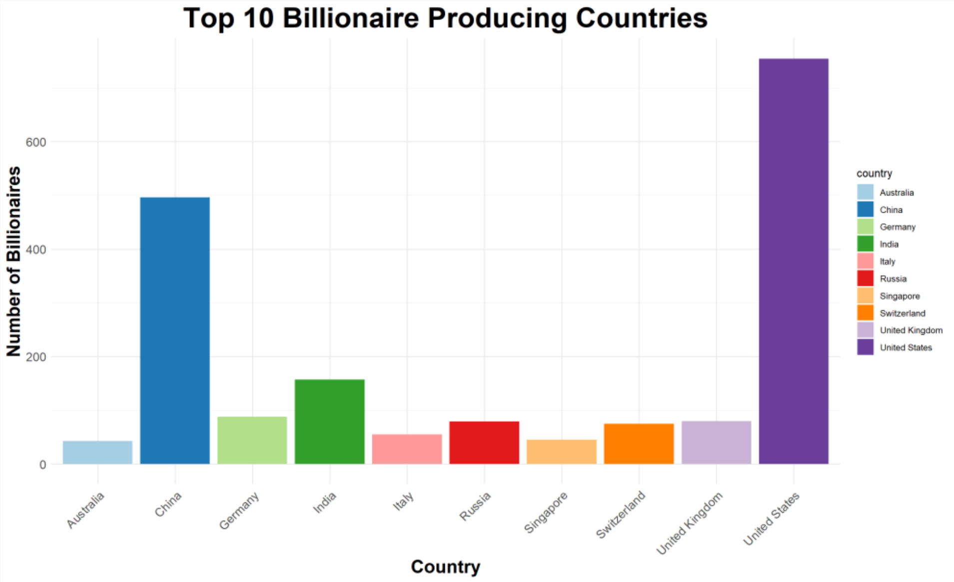
VARIABLE SUMMARIES AND CORRELATION



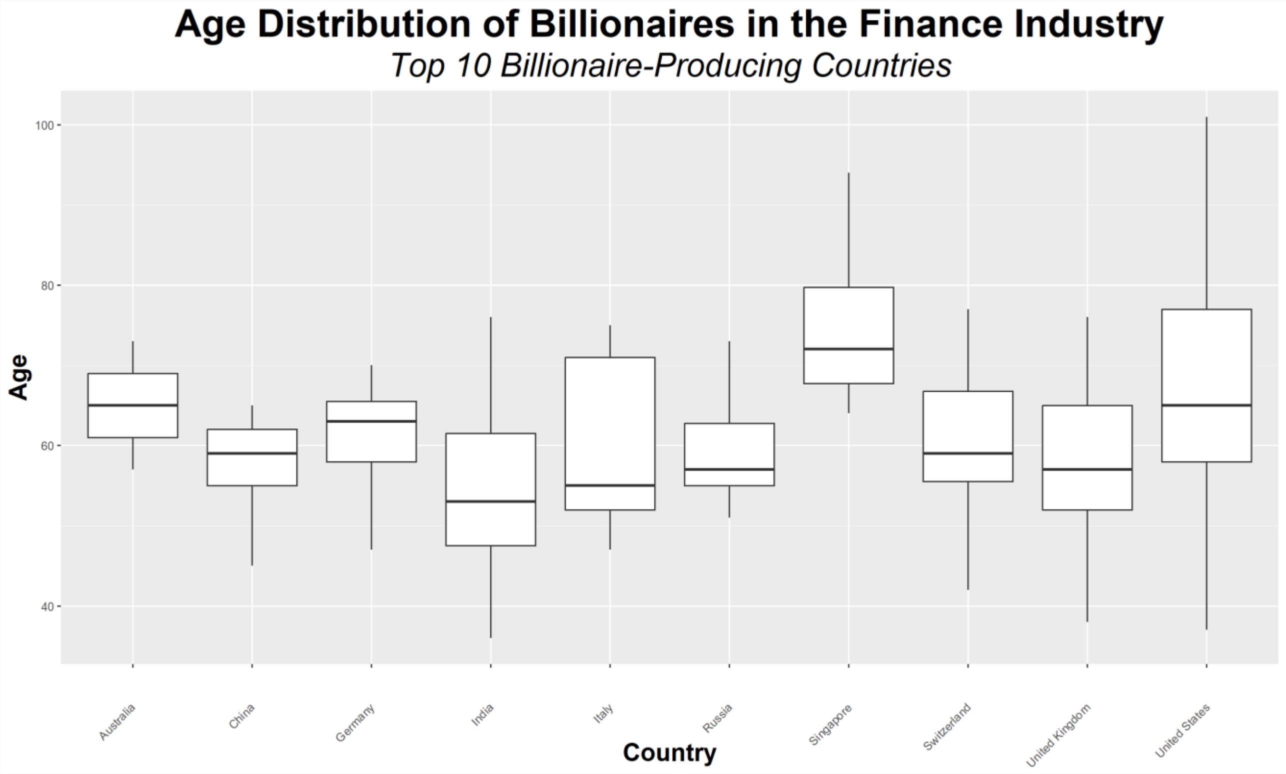
WHAT IS THE TOP BILLIONAIRE PRODUCING INDUSTRY, AND BASED ON THIS ANSWER, WHAT ARE THE AGE DISTRIBUTIONS OF THIS INDUSTRY WITHIN THE TOP 10 BILLIONAIRE PRODUCING COUNTRIES?

TOP 5 INDUSTRIES AND TOP 10 COUNTRIES





AGE DISTRIBUTION OF BILLIONAIRES IN FINANCE INDUSTRY



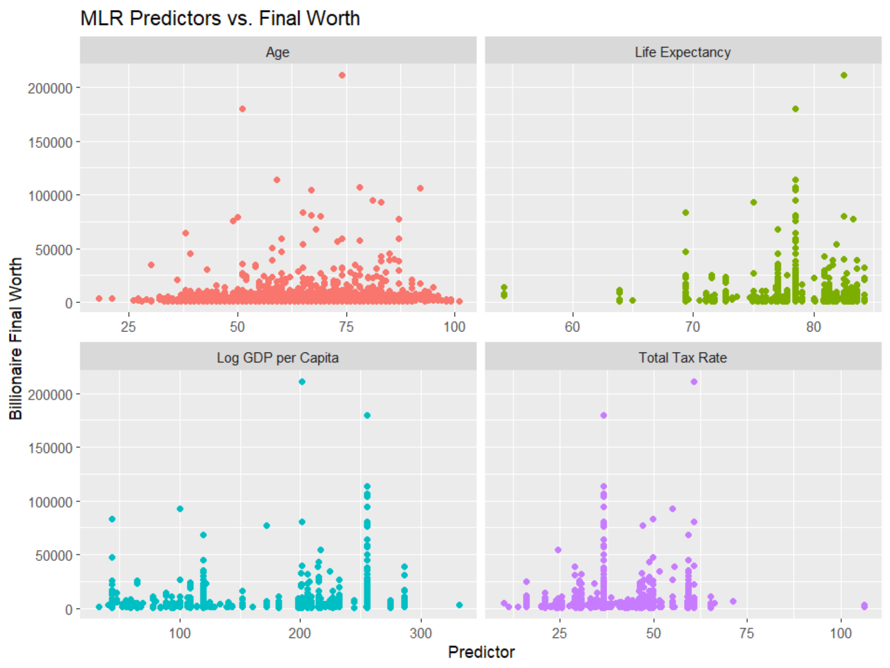
BASED ON PERSONAL CHARACTERISTICS AND COUNTRY'S ECONOMIC FACTORS, CAN WE PREDICT A BILLIONAIRE'S FINAL WORTH?

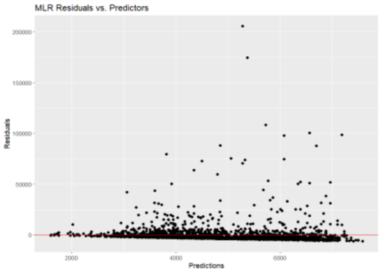
DATA VISUALIZATION – MLR I

MULTIPLE LINEAR REGRESSION I

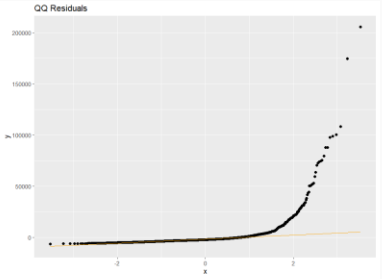
PARAMETER SELECTION:

- LEAST SQUARES WITH STEP AIC
- VARIABLE TRANSFORMATIONS
- CATEGORICAL PREDICTORS





ASSUMPTIONS FAILED

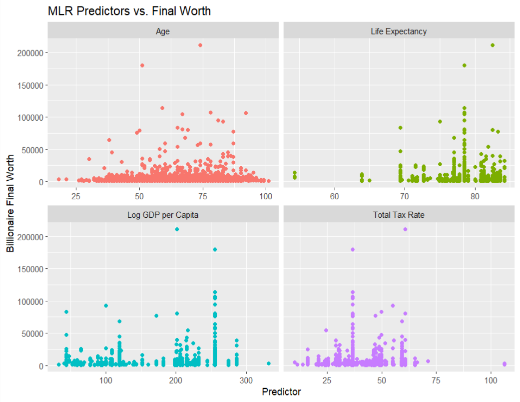


MULTIPLE LINEAR REGRESSION I

MODEL PREDICTORS:

- AGE
- TOTAL TAX RATE
- LIFE EXPECTANCY
- LOG GDP PER CAPITA

CREATED FROM EXISTING DATA



- R-SQUARED: 0.013
- RMSE: 9985.89 (9.98 BILLION)
  - USING A 5-FOLD CV
- ALL VIFS UNDER 5
  - NO MULTICOLLINEARITY

★ MOST SIGNIFICANT PREDICTOR: LOG GDP PER CAPITA

MULTIPLE LINEAR REGRESSION I

MODEL PREDICTORS:
- AGE
- TOTAL TAX RATE
- LIFE EXPECTANCY
- LOG GDP PER CAPITA

CREATED FROM EXISTING DATA

MOST SIGNIFICANT PREDICTOR:
LOG GDP PER CAPITA

OPRAH WINFREY: $2.5 BILLION ACTUAL WORTH
- PREDICTION: $5.81 BILLION
- CI: (5.12, 6.51)

GIORGIO ARMANI: $11.1 BILLION ACTUAL
- PREDICTION: $3.38 BILLION
- CI: (2.00, 4.76)

HANNAH VALENTY: ?? BILLION ACTUAL
- PREDICTION: $4.09 BILLION
- PI: (-15.88, 24.06)

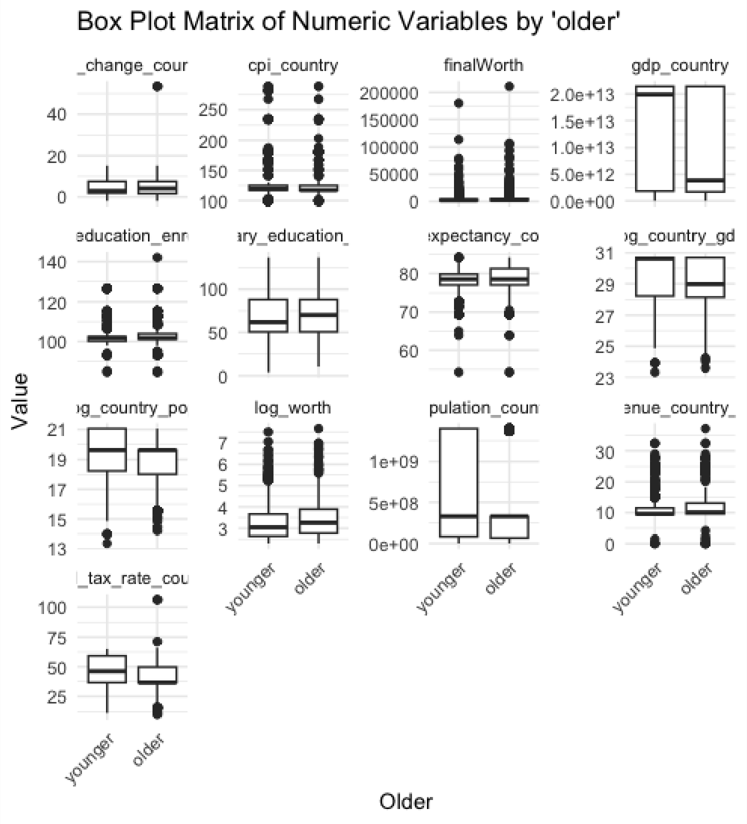
CONCLUSIONS – MLR I

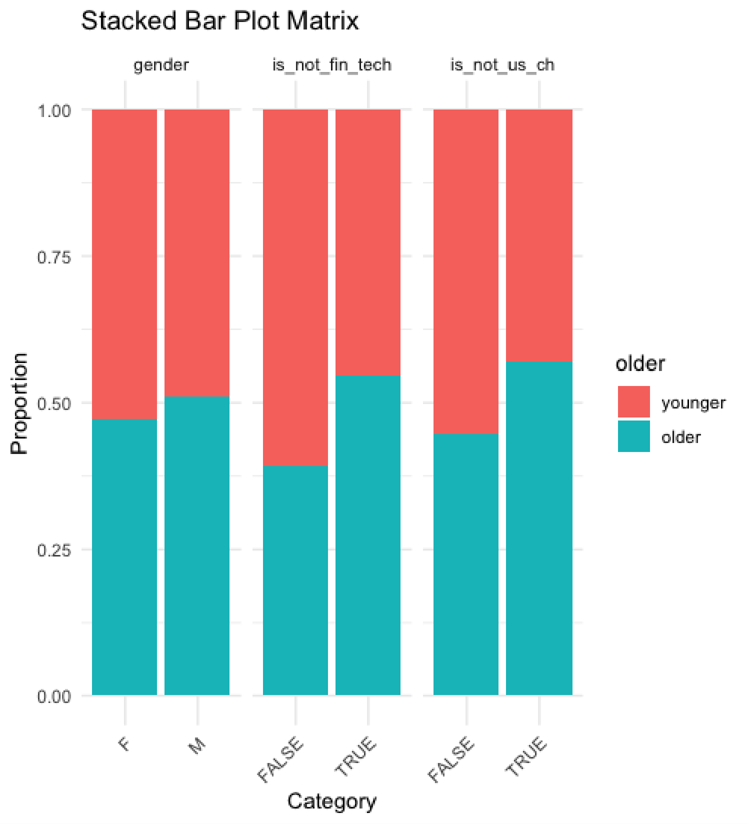
- While this model was not the best predictor, it guided us to create the GDP per capita feature
- Perhaps a future question about final worth could be asked using data about more localized regions, removing the constraint of repeated country values

FOR BOTH SELF-MADE AND NON-SELF-MADE BILLIONAIRES, WHICH FACTORS SEPARATE OLDER BILLIONAIRES FROM YOUNGER ONES?

- LOGISTIC REGRESSION -

VARIABLE EXPLORATION- LOGISTIC REGRESSION





MODEL I - SELF MADE BILLIONAIRES

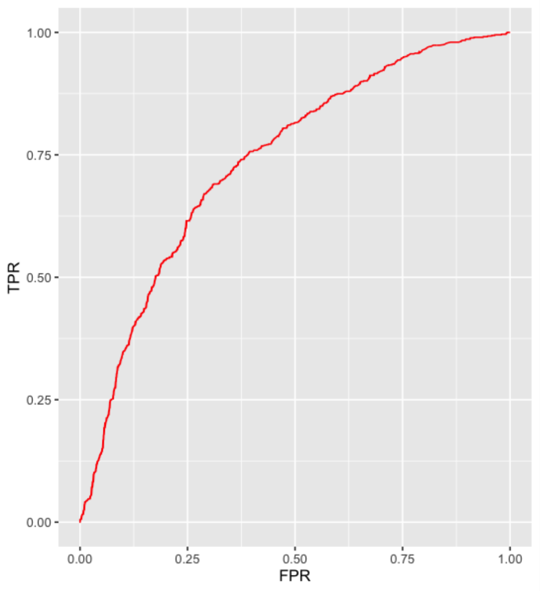
```
summary(sm_logit_model3_lasso)
```

```
Call:
glm(formula = older ~ is_not_fin_tech + is_not_us_ch + log_worth +
    cpi_country + cpi_change_country + tax_revenue_country_country +
    total_tax_rate_country, family = "binomial", data = sm_logit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0234  -0.9875  -0.4730   0.9514   2.5790

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  0.767322   0.470619   1.630 0.103006
is_not_fin_techTRUE          1.065439   0.125318   8.502  < 2e-16 ***
is_not_us_chTRUE             2.824790   0.270516  10.442  < 2e-16 ***
log_worth                    0.233976   0.066613   3.512 0.000444 ***
cpi_country                 -0.043595   0.004521  -9.644  < 2e-16 ***
cpi_change_country           0.412706   0.036817  11.210  < 2e-16 ***
tax_revenue_country_country -0.049606   0.014958  -3.316 0.000912 ***
total_tax_rate_country       0.018539   0.007047   2.631 0.008517 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2341.4  on 1692  degrees of freedom
Residual deviance: 2032.1  on 1685  degrees of freedom
AIC: 2048.1

Number of Fisher Scoring iterations: 4
```

```{r}
vif(sm_logit_model3_lasso)
```

```
            is_not_fin_tech                is_not_us_ch                   log_worth                 cpi_country
                   1.218380                    6.148328                    1.018880                    3.691705
         cpi_change_country tax_revenue_country_country      total_tax_rate_country
                   3.932018                    2.068226                    2.461667
```

Area under the curve: 0.7366



MODEL I EVALUATION- LOGISTIC REGRESSION


```
table(predictions$pred, predictions$obs)
```

```
          younger older
  younger     665   299
  older       230   499
```

```
# model predicts Mark Cuban correctly
mark_cuban<-sm_logit[592,c(29,28,24,15,16,21,22)]

predict(sm_logit_model3, mark_cuban, type="response")
```

```
     592
0.696402
```

```
Accuracy: 0.688
True Positive Rate (TPR): 0.743
True Negative Rate (TNR): 0.625
False Positive Rate (FPR): 0.287
False Negative Rate (FNR): 0.257
```

MODEL 2 - NON SELF-MADE BILLIONAIRES

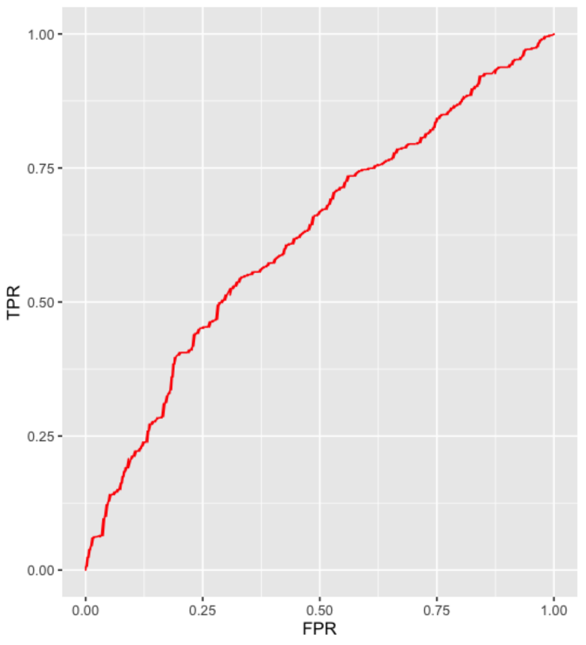
```
summary(nsm_logit_model)
```

```
Call:
glm(formula = older ~ gender + is_not_fin_tech + finalWorth +
    total_tax_rate_country + population_country, family = "binomial",
    data = nsm_logit)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.993  -1.226   0.794   1.023   1.669

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.803e-01  3.770e-01   0.478 0.632341
genderM                 3.570e-01  1.727e-01   2.067 0.038714 *
is_not_fin_techTRUE     6.051e-01  2.265e-01   2.672 0.007551 **
finalWorth              2.830e-05  1.211e-05   2.337 0.019460 *
total_tax_rate_country -2.309e-02  7.417e-03  -3.113 0.001854 **
population_country      6.918e-10  1.929e-10   3.586 0.000336 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 968.17  on 713  degrees of freedom
Residual deviance: 929.03  on 708  degrees of freedom
AIC: 941.03

Number of Fisher Scoring iterations: 4
```

```{r}
vif(nsm_logit_model)
```

```
     gender    is_not_fin_tech    finalWorth total_tax_rate_country    population_country
   1.013108           1.008911      1.005882               1.102172              1.104887
```

Area under the curve: 0.6223



MODEL 2 EVALUATION- LOGISTIC REGRESSION

```
# model predicts Trudy Cathy White correctly
trudy_cathy_white<-sm_logit[126,c(1,11,29,22,23)]

predict(nsm_logit_model, trudy_cathy_white, type="response")
```

```
      126
0.7338769
```

	Actual	
Predicted	younger	older
younger	149	135
older	146	284

```
Accuracy: 0.606
True Positive Rate (TPR): 0.678
True Negative Rate (TNR): 0.505
False Positive Rate (FPR): 0.495
False Negative Rate (FNR): 0.322
```

** Balance between overfitting and low accuracy**

CONCLUSIONS - LOGISTIC REGRESSION

1. For self-made billionaires, economic data about their countries is important in predicting their age (compared to non-self-made billionaires)

2. Industry is an important predictor for both
   - Billionaires in the tech and finance industries are less likely to be older

3. Financial worth is an important predictor for both
   - Richer billionaires are more likely to be older

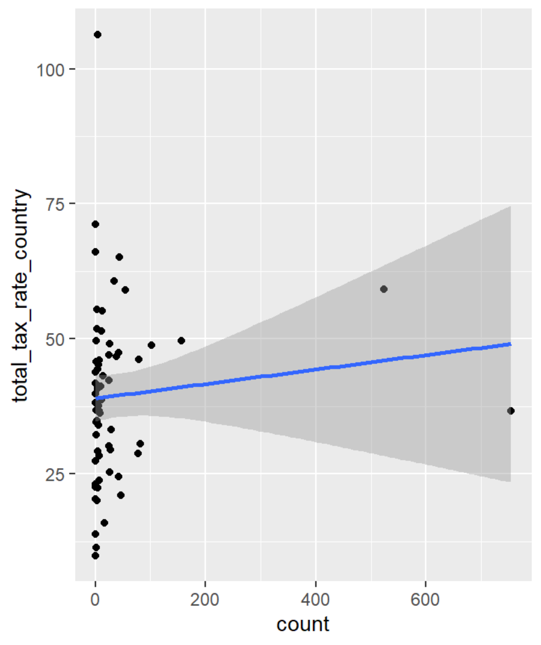
AGGREGATION: WHAT ECONOMIC FACTORS LEAD TO A COUNTRY HAVING A LOT OF BILLIONAIRES?

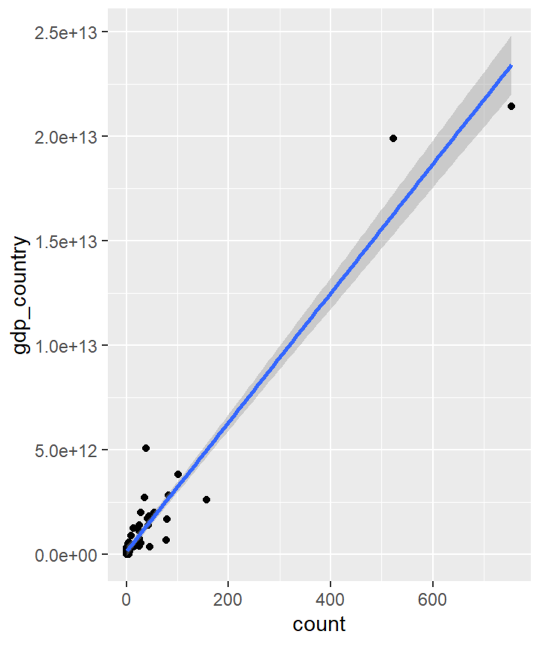
PRINCIPAL COMPONENT ANALYSIS – DATA EXPLORATION

75 COUNTRIES IN THE DATASET, BUT ONLY 64 HAVE COUNTRY STATISTICS

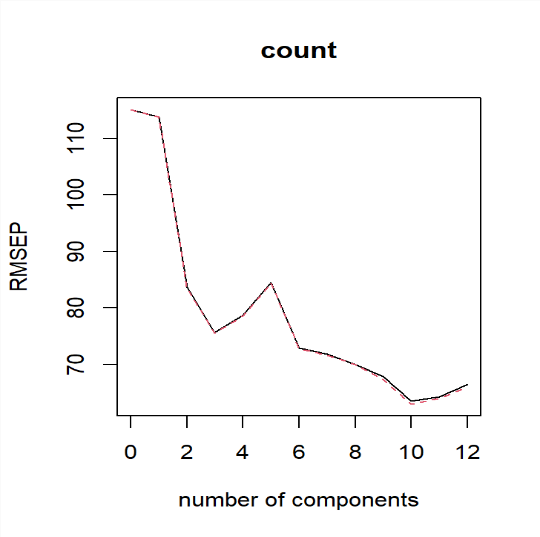
NOTABLE COUNTRIES MISSING STATISTICS: HONG KONG, MONACO, TAIWAN

THE PROBLEM OF SUCH A SMALL DATASET LED US TO PCR



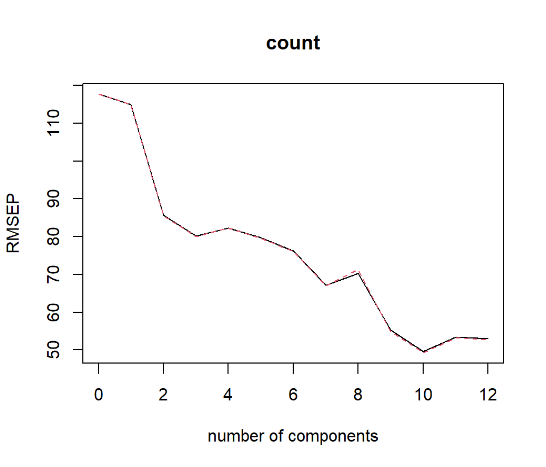


PRINCIPAL COMPONENT ANALYSIS – MODEL SELECTION



MODEL 1: ALL 64 COUNTRIES

- MAE (10 COMPONENTS): 15.55
- RMSE (10 COMPONENTS): 24.06



MODEL 2: REMOVED COUNTRIES WITH OUTLIER CPI STATISTICS (ARGENTINA, NIGERIA, EGYPT)

- MAE (10 COMPONENTS): 16.77
- RMSE (10 COMPONENTS): 22.70

PRINCIPAL COMPONENT ANALYSIS - PREDICTION

	country	count	prediction	residual	cpi_country	gdp_country	total_tax_rate_country	population_country
1	Australia	43	38	5	119.80	1.392681e+12	47.4	25766605
2	Bahrain	1	21	-20	117.59	3.857407e+10	13.8	1501635
3	China	523	572	-49	125.08	1.991000e+13	59.2	1397715000
4	Colombia	1	3	-2	140.95	3.238028e+11	71.2	50339443
5	India	157	113	44	180.44	2.611000e+12	49.7	1366417754
6	Japan	38	123	-85	105.48	5.081770e+12	46.7	126226568
7	New Zealand	2	2	0	114.24	2.069288e+11	34.6	4841000
8	Qatar	2	-12	14	115.38	1.834662e+11	11.3	2832067
9	Russia	79	56	23	180.75	1.699877e+12	46.2	144373535
10	Singapore	46	14	32	114.41	3.720625e+11	21.0	5703569
11	Spain	25	30	-5	110.96	1.394116e+12	47.0	47076781
12	Sweden	26	32	-6	110.51	5.308329e+11	49.1	10285453
13	United Kingdom	82	75	7	119.62	2.827113e+12	30.6	66834405
14	United States	754	687	67	117.24	2.142770e+13	36.6	328239523

- SEEMS TO PREDICT TRADITIONALLY FIRST-WORLD COUNTRIES REASONABLY WELL, WITH SOME EXCEPTIONS
- POTENTIAL TO CONSTRAIN THE MODEL TO SOME EXTENT – NEGATIVE NUMBERS DON'T MAKE SENSE
- WHY ARE BAHRAIN AND QATAR SO DIFFERENT?
- WHY DOES THE MODEL DO SUCH A POOR JOB PREDICTING JAPAN AND SINGAPORE? POTENTIAL CLUES TO BE FOUND IN A MULTILINEAR REGRESSION MODEL?

LIMITATIONS AND FUTURE WORK

- Some assumptions were not met to be able to create accurate models
- More data about the billionaires and the source of their wealth
- More data about the wealth distributions and demographics in the countries
- If the data was available, we could investigate at what age people became a billionaire and see if including that would change our models