**Report Summary and Outputs**

Research question: Does the region of origin influence the income of immigrants compared with non-immigrants in Canada, and if so why?

A few thoughts, the research questions did not clearly suggest that some recategorisation will be needed such as education. Income was a continuous variable so a linear model would have been used instead of binarising the variable. We don’t know what informed the choice of the cutoff point. Also, the income variable in the dataset excluded people below 16 years, as expected. I think the region of origin was not eventually captured in the analysis.

Analysis/codes: A few variable names in the script did not match what they were in the original dataset (Eg. Sex, TotInc), leading to error messages. Attempting to select only the needed variables from the dataset using the code led to a severe alteration in the data frame with a drastic number of missing data. Other than this the codes ran successfully and gave the tables below.

# A tibble: 4 × 3

POB\_Binary ImmigrationStat count

*<dbl>* *<dbl>* *<int>*

1 0 0 539052

2 0 1 19

3 1 0 2754

4 1 1 194449

Call:

glm(formula = topearner ~ POB\_Binary + ImmigrationStat + Sex +

DegreeEarned + PR + AGEIMM, family = "binomial", data = census\_clean)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.3303845 0.0453738 -51.360 <2e-16 \*\*\*

POB\_Binary -0.0141918 0.0518599 -0.274 0.784

ImmigrationStat -1.4471692 0.0652471 -22.180 <2e-16 \*\*\*

Sex 0.8226602 0.0069096 119.060 <2e-16 \*\*\*

DegreeEarned 0.0084512 0.0002841 29.749 <2e-16 \*\*\*

PR 0.0146957 0.0002531 58.063 <2e-16 \*\*\*

AGEIMM -0.0119612 0.0004339 -27.567 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 616520 on 736273 degrees of freedom

Residual deviance: 594803 on 736267 degrees of freedom

AIC: 594817

Number of Fisher Scoring iterations: 5