Q 5 Salary Data

Ana Paula Felix de Queiroz

Question 5:

- Are specific job titles being offered to candidates of a specific nationality?

- Does the nationality of the candidate affect the status of case?

- Does job subtitle of the candidate affect the status o the case?

Data pre processing

It is important to notice that for this analysis we had a significant amount of missing values values on the column of interest. for instance we have a data set with 167278 observations but for column country of citizenship we have a total of 124106 missing values. Nevertheless, we will proceed with the analysis using the remaining 43172 observations.
Since we are focusing our analysis on data related jobs we are also excluding all the job subtitles unrelated to data, as well

```
## The following objects are masked from df:
##
##     CASE_NUMBER, CASE_RECEIVED_DATE, CASE_STATUS,
##     COLLEGE_MAJOR_REQUIRED, COUNTRY_OF_CITIZENSHIP, DECISION_DATE,
##     EDUCATION_LEVEL_REQUIRED, EMPLOYER_NAME,
##     EXPERIENCE_REQUIRED_NUM_MONTHS, EXPERIENCE_REQUIRED_Y_N,
##     FULL_TIME_POSITION_Y_N, JOB_TITLE, JOB_TITLE_SUBGROUP, order,
##     PAID_WAGE_PER_YEAR, PAID_WAGE_SUBMITTED, PAID_WAGE_SUBMITTED_UNIT,
##     PREVAILING_WAGE_PER_YEAR, PREVAILING_WAGE_SOC_CODE,
##     PREVAILING_WAGE_SOC_TITLE, PREVAILING_WAGE_SUBMITTED,
##     PREVAILING_WAGE_SUBMITTED_UNIT, VISA_CLASS, WORK_CITY,
##     WORK_POSTAL_CODE, WORK_STATE, WORK_STATE_ABBREVIATION
```

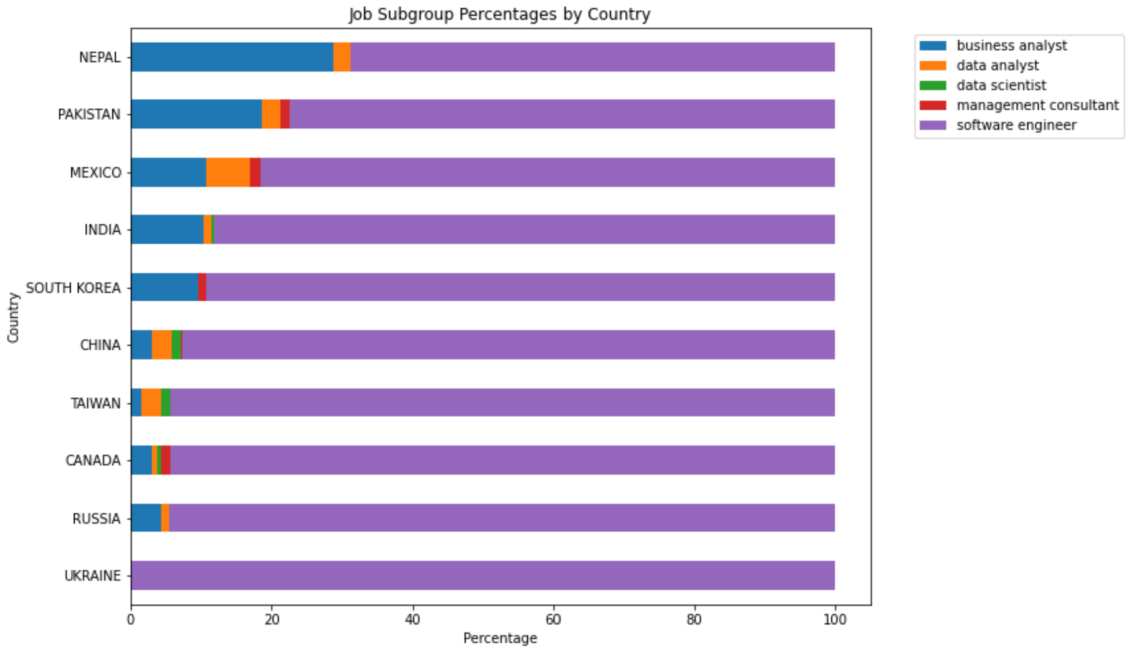
```
## [1] 133012     28
```

- Are specific job titles being offered to candidates of a specific nationality? In other words, we want to know if there a significant difference in number of hiring for a particular job subtitle considering country of citizenship.
First we conducted an analysis in python looking at the top10 countries with higher hiring frequency. We then looked at the frequency of job subtitle in each of these countries.

JOB_TITLE_SUBGROUP COUNTRY_OF_CITIZENSHIP	business analyst	data analyst	data scientist	management consultant	software engineer
INDIA	10.433286	1.094948	0.297200	0.109495	88.065071
CHINA	3.100775	2.768549	1.328904	0.110742	92.691030
CANADA	3.004292	0.858369	0.429185	1.287554	94.420601
SOUTH KOREA	9.677419	0.000000	0.000000	1.075269	89.247312
RUSSIA	4.347826	1.086957	0.000000	0.000000	94.565217
NEPAL	28.750000	2.500000	0.000000	0.000000	68.750000
PAKISTAN	18.666667	2.666667	0.000000	1.333333	77.333333
TAIWAN	1.428571	2.857143	1.428571	0.000000	94.285714
UKRAINE	0.000000	0.000000	0.000000	0.000000	100.000000
MEXICO	10.769231	6.153846	0.000000	1.538462	81.538462

Key takeaways considering the top 10 countries with higher hiring percentage:

- Software engineers are the most sough out professionals across the board in the top 10 countries.
- Countries such as Canada, Russia, Taiwan, and China have over 90% of its hiring pool being directed to software engineers professionals.
- Ukraine only hire software engineers
- Nepal has the highest percentage of individuals working as business analysts, with over 28%. followed by Pakistan with over 18%.
- Mexico has the highest percentage of individuals working as data analysts with over 6% of the pool of job titles.
- Taiwan and China have the highest percentage of
- China and Taiwan have the highest percentage of individuals working as data scientists, with over 1.4% and 1.3%.
- Pakistan has the highest percentage of individuals working as management consultants, with over 1.3%.



- Does the nationality of the candidate affect the status of case? We wanted to investigate whether the nationality of the candidate affects the status of the case (positive or negative) in the context of immigration applications.

CASE_STATUS	certified	certified-expired	denied	withdrawn
COUNTRY_OF_CITIZENSHIP				
INDIA	4158	1864	279	286
CHINA	670	340	34	57
CANADA	236	151	19	19
SOUTH KOREA	154	82	14	19
RUSSIA	70	30	6	3
NEPAL	55	28	6	3
PAKISTAN	50	26	4	4
TAIWAN	72	33	3	2
UKRAINE	45	20	2	4
MEXICO	124	65	58	8

The following tests aimed to investigate the relationship between nationality and case status (positive or negative) To achieve this, we categorized certified withdraw, certified-expired, and certified as positive, and denied as negative. We employed logistic regression analysis and ANOVA to assess the significance of nationality as a predictor of case status.

Our logistic regression analysis showed a significant relationship ($p < 0.001$) between the predictor variable (country) and the response variable (positive or negative status), as indicated by the significant intercept.

However, upon analyzing the individual coefficients for different countries, not all of them were found to be significant ($p > 0.05$).

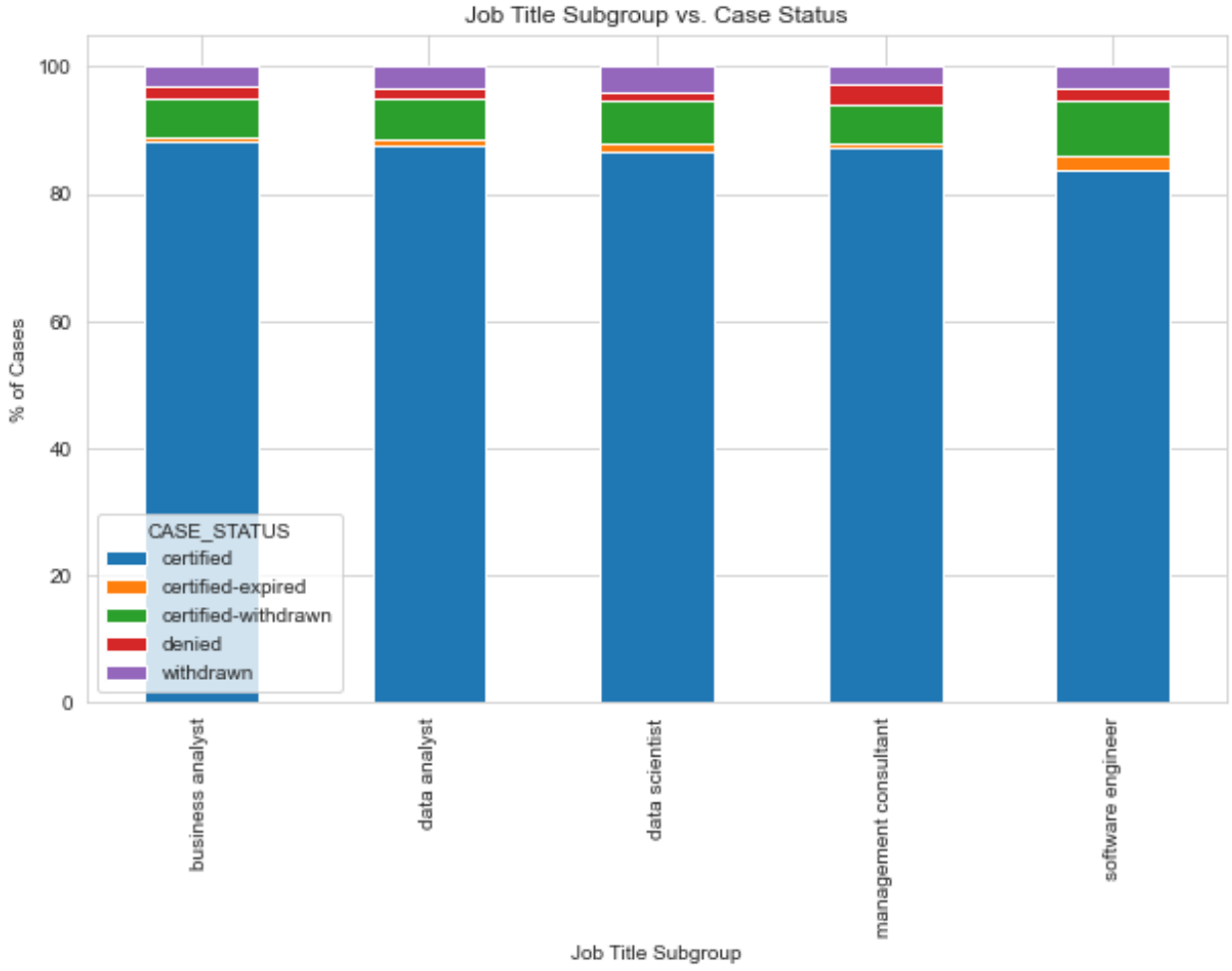
Given the sparse nature of the data and the non-conformity of the model assumptions, we ran a Monte Carlo simulation and a Fisher test. Both these tests yielded non-significant results. Therefore, we can conclude that the nationality of the candidate does not influence the case status in this dataset. Nonetheless, it is essential to consider other factors that may affect the case status in actual cases.

```
## Warning in chisq.test(status[, -1]): Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  status[, -1]
## X-squared = 31.17, df = 27, p-value = 0.2642

## Warning: package 'brglm2' was built under R version 4.2.3

##
## Call:
## glm(formula = cbind(Positive, Negative) ~ factor(country), family = binomial,
##     data = new_status)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 3.10459    0.32324   9.605   <2e-16 ***
## factor(country)CHINA        0.58543    0.38868   1.506    0.132
## factor(country)INDIA        0.03631    0.32927   0.110    0.912
## factor(country)MEXICO       1.05430    1.05835   0.996    0.319
## factor(country)NEPAL       -0.59228    0.53354  -1.110    0.267
## factor(country)PAKISTAN    -0.22820    0.60710  -0.376    0.707
## factor(country)RUSSIA      -0.01354    0.60485  -0.022    0.982
## factor(country)SOUTH KOREA  0.71313    0.78452   0.909    0.363
## factor(country)TAIWAN       1.12952    1.05782   1.068    0.286
## factor(country)UKRAINE      1.06980    1.05824   1.011    0.312
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.4580e+01  on 9  degrees of freedom
## Residual deviance: 8.2157e-15  on 0  degrees of freedom
## AIC: 54.96
##
## Number of Fisher Scoring iterations: 4

## Analysis of Deviance Table (Type II tests)
##
## Response: cbind(Positive, Negative)
##                  LR Chisq Df Pr(>Chisq)
## factor(country)    14.58  9     0.1031
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  10000 replicates)
##
## data:  cont_table
## p-value = 0.1524
## alternative hypothesis: two.sided
```

Does job subtitle of the candidate affect the status o the case? To begin, we used Python to create a stacked bar chart to visualize the relationship between job title subgroup and case status. The chart indicated that there may not be a significant difference in case status frequency between job title subgroups. However, to confirm this, we conducted a chi-square test for independence.

The results of the chi-square test suggested that there is a significant association between job title subgroup and case status. The extremely low p-value ($< 2.2e\text{-}16$) indicated that the null hypothesis of independence between the two variables can be rejected. Therefore, we can conclude that the job title subgroup of the candidate does indeed have an effect on the case status.




```
## [1] "There is a significant association between job title subgroup and case status."
```

To further explore this, we created a mosaic graph, which provided a more detailed understanding of the relationship between job title subgroup and case status. We found that business analysts are more likely to be

certified compared to software engineers, and are less likely to have their certification expired or withdrawn. Data analysts are less likely to have their certification expired or withdrawn compared to software engineers.

On the other hand, software engineers are more likely to have a denied visa and are less likely to have a certified case status compared to business analysts. They are also more likely to have their certification expired, which suggests that companies didn't move forward with the hiring after their case was certified, as well as have their certification withdrawn, indicating that the application was withdrawn before the certification was granted. In conclusion, the job title subgroup of the candidate is a significant factor affecting the status of their case, and this information could be useful for companies in their hiring process.

```
## Loading required package: grid
```